lume XIII *July 1959* *Number 67*

Mathematical Tables *and other* Aids to Computation

Published Quarterly
by the
National Academy of Sciences—National Research Council

Information to Subscribers

The journal is published quarterly in one volume per year with issues numbers serially since Volume I, Number 1. Starting with January, 1959 subscriptions are $8.00 per year, single copies $2.50. Back issues are available as follows:

Volume I (1943–1945), Nos. 10 and 12 *only* are available; $1.00 per issue.
Volume II (1946–1947), Nos. 13, 14, 17, 18, 19, and 20 *only* available; $1.00 per issue.
Volume III (1948–1949), Nos. 21–28 available. $4.00 per year (four issues), $1.25 per issue.
Volume IV (1950 through 1958), all issues available; $5.00 per year, $1.50 per issue.

Agents for Great Britain and Ireland: Scientific Computing Service, Ltd.
Bedford Square, London W. C. 1

Microcard Edition

Volumes I–X (1943–1956), Nos. 1–56 are now available on Microcards and may be purchased from the Microcard Foundation, Box 2145, Madison 5, Wisconsin, at a cost of $20.00 for the complete set. Future volumes will be available on Microcards in the year following original publication.

Information to Contributors

All contributions intended for publication in *Mathematical Tables and Other Aids to Computation* and all books for review should be addressed to H. Polachek, Technical Director, Applied Mathematics Laboratory, David Taylor Model Basin, Washington 7, D. C. The author may suggest an appropriate editor for his paper.

Subscriptions, address changes, business communications and payments should be sent to:

The Printing and Publishing Office
The National Academy of Sciences
2101 Constitution Avenue
Washington 25, D. C.

PUBLISHED BY
THE NATIONAL ACADEMY OF SCIENCES—NATIONAL RESEARCH COUNCIL
Mt. Royal and Guilford Avenues, Baltimore 2, Md.
Second-class postage paid at Baltimore, Md.

Ma
equ
pre
wit
tati
trat
T
von
and
syst
we
refe
Som
wor
the

2
the
vert
velo
$u(x,$
the
Th
to th

(2.1

(2.2)

(2.3)

(2.4)

Equa
is th
of m

Re
paper
* F

A Study of a Numerical Solution to a Two-Dimensional Hydrodynamical Problem

By A. Blair, N. Metropolis, J. von Neumann,* A. H. Taub & M. Tsingou

1. Introduction. The purpose of this paper is to report the results obtained on Maniac I when that machine was used to solve numerically a set of difference equations approximating the equations of two-dimensional motion of an incompressible fluid in Eulerian coordinates. More precisely, the problem was concerned with the two-dimensional motion of two incompressible fluids subject only to gravitational and hydrodynamical forces which at time $t = 0$ were distributed as illustrated in Fig. 1.

This problem was discussed and formulated for machine computation by John von Neumann and others. His own original draft of a discussion of the differential and difference equations is given in Appendix I, and an iteration scheme for solving systems of linear equations is given in Appendix II. In the main body of this paper we shall outline the derivation of the equations employed by the computer and refer to these appendices for detailed discussions concerning them where necessary. Some of von Neumann's difference equations were modified in the course of the work. The reasons for these modifications and their nature will be enlarged upon in the course of the discussion.

2. The Equations of Motion and Boundary Conditions. We denote by x and y the Cartesian abscissa and ordinate of a point in a fixed coordinate system in a vertical plane oriented as in Fig. 1; that is, x and y are Eulerian coordinates. The velocity of the fluid at this point at time t will be said to have x and y components $u(x, y, t)$ and $v(x, y, t)$, respectively. The density of the fluid will be denoted by ρ, the pressure by p, and the acceleration of gravity by g.

The system of equations describing the motion of an incompressible fluid subject to the force of gravity in the vertical direction is then

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{1}{\rho}\frac{\partial p}{\partial x} \tag{2.1}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} = -\frac{1}{\rho}\frac{\partial p}{\partial y} + g \tag{2.2}$$

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + v\frac{\partial \rho}{\partial y} = 0 \tag{2.3}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{2.4}$$

Equations (2.1) and (2.2) represent the conservation of momentum, equation (2.3) is the incompressibility condition, and equation (2.4) states the conservation of mass.

Received April 19, 1959. This work was done at Los Alamos Scientific Laboratory and the paper is a condensation of Report LA-2165 with the same title.

* Published posthumously.

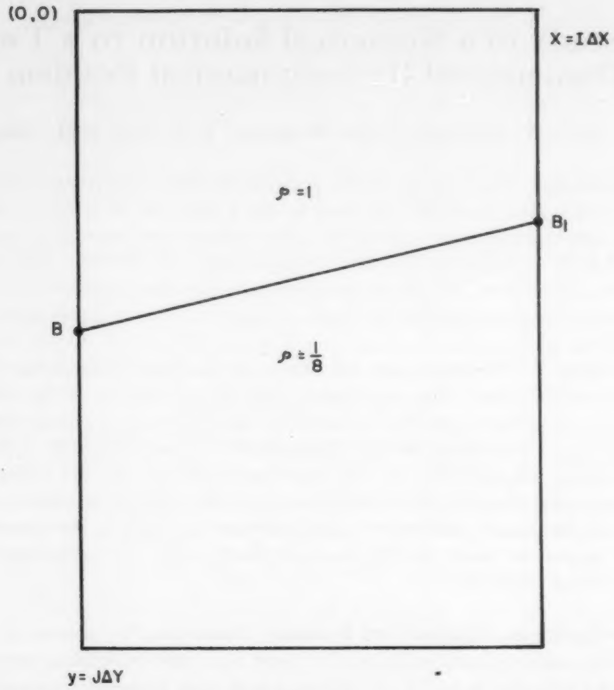

FIG. 1.—Initial density distribution. BB_1 denotes the boundary separating the fluid of density $\rho = 1$ from that of density $\rho = \frac{1}{8}$ at time $t = 0$.

At any exterior boundary of the fluid the component of velocity normal to the boundary vanishes. That is

$$u \cos(x, n) + v \cos(y, n) = 0 \tag{2.5}$$

where $\cos(x, n)$ and $\cos(y, n)$ are the cosines of the angles between the x axis and the normal to the boundary and the y axis and the normal to the boundary, respectively.

Along a curve across which there is a density discontinuity we must have the component of the velocity normal to the curve continuous. That is, we must have

$$[u] \cos(x, n) + [v] \cos(y, n) = 0 \tag{2.6}$$

where

$$[f] = f(x^+, y^+) - f(x^-, y^-)$$

and x^+, y^+ and x^-, y^- represent contiguous points on opposite sides of the curve of discontinuities.

3. Discontinuities. The only discontinuities present in incompressible fluid motion are density discontinuities and across these equation (2.6) must hold. In the calculation to be described, such discontinuities are not explicitly taken into account. Indeed, it was one purpose of the calculation to see if this type of discontinuity could be followed in time from a plot of the density contours when no special provision was made to provide for the discontinuities.

Initially the density distribution assumed was that given in Fig. 1. Some provisions must be made in the difference equations to represent the spatial derivatives of the density on the curve BB_1 at time $t = 0$ (and at subsequent times on the curve into which BB_1 moves). We shall discuss this point subsequently.

4. The Stream Function. It follows from equation (2.4) that there exists a function ψ called the stream function such that

$$u = -\psi_y, \qquad v = \psi_x \tag{4.1}$$

On an exterior boundary, equation (2.5) obtains and this may be written as

$$-\psi_y \cos(x, n) + \psi_x \cos(y, n) = 0 \tag{4.2}$$

If the equation of the boundary is given parametrically by

$$x = x(s), \qquad y = y(s)$$

with

$$\left(\frac{dx}{ds}\right)^2 + \left(\frac{dy}{ds}\right)^2 = 1$$

then the direction cosines of the normal are

$$\cos(x, n) = -\frac{dy}{ds}$$

$$\cos(y, n) = \frac{dx}{ds}$$

and equation (4.2) becomes

$$\psi_x \frac{dx}{ds} + \psi_y \frac{dy}{ds} = 0$$

That is

$$\psi = \text{constant}$$

on the exterior boundary. This constant may be chosen to be zero and the exterior boundary condition becomes

$$\psi = 0 \tag{4.3}$$

5. Equations Determining the Stream Function and Density. Substituting from equation (4.1) into equations (2.1) and (2.2), we obtain

$$\rho(\psi_{ty} - \psi_y\psi_{xy} + \psi_x\psi_{yy}) = p_x$$

$$\rho(\psi_{tx} - \psi_y\psi_{xx} + \psi_x\psi_{xy}) = -p_y + g\rho$$

where the subscript denotes a derivative with respect to the variable indicated. Differentiating the first of these with respect to y and the second with respect to x and adding, we have

$$(\rho\psi_{tx})_x + (\rho\psi_{ty})_y = \rho_x(g - {}^1I) - \rho_y{}^2I - \rho{}^3I = -\omega \tag{5.1}$$

where

$${}^1I = \psi_x\psi_{xy} - \psi_y\psi_{xx} \tag{5.2}$$

$${}^2I = \psi_x\psi_{yy} - \psi_y\psi_{xy} \tag{5.3}$$

$${}^3I = \psi_x\lambda_y - \lambda_x\psi_y \tag{5.4}$$

with

$$\lambda = \psi_{xx} + \psi_{yy} \tag{5.5}$$

If we now set

$$\chi = -\psi_t \tag{5.6}$$

then

$$-(\rho\chi_x)_x - (\rho\chi_y)_y = -\omega \tag{5.7}$$

where ω is given in terms of ψ and ρ by the right-hand side of equation (5.1), and on the boundary

$$\chi = 0 \tag{5.8}$$

We may regard equation (5.1) and the boundary condition $\psi = 0$ as equations for determining ψ and use equation (4.1) to define u. The pressure may then be calculated from equations (2.1) or (2.2).

The density may then be determined from equation (2.3), which may be written as

$$\rho_t = \psi_y\rho_x - \psi_x\rho_y \tag{5.9}$$

The system of equations (5.6) through (5.9) defines the problem to be approximated by difference equations and to be solved numerically on the computer.

6. The Difference Equations. For any function $a(x, y, t)$ let

$$a^h_{i,j} = a(x_i, y_j, t^h) \tag{6.1}$$

where

$$\begin{array}{lll} & x_i \equiv i\Delta x & i = 0, 1, \cdots I - 1, I \\ (6.2) & y_j \equiv j\Delta y & j = 0, 1, \cdots J - 1, J \\ & t^h \equiv h\Delta t & h = 0, 1, 2, \cdots \end{array}$$

and I and J are integers. Occasionally indices $i \pm \frac{1}{2}$, $j \pm \frac{1}{2}$, and $h + \frac{1}{2}$ are used.

The difference equations we shall consider will involve quanties $\psi^h_{i,j}$, $\chi^h_{i,j}$, and $\rho^h_{i,j}$ for h, i, and j given by equation (6.2). The mesh points with $i = 0$ or I $(j = 0, 1, \cdots J)$ or $j = 0$ or J $(i = 0, 1, \cdots I)$ are said to be boundary points.

The remaining mesh points are called interior points. Equations (4.3) and (5.8) define $\psi^h_{i,j} = \chi^h_{i,j} = 0$ for boundary points. Hence, we must give an algorithm for determining $\psi^h_{i,j}$ for interior points and $\rho^h_{i,j}$ for interior and boundary points. This algorithm involves the finite difference representation of equation (5.7) for interior points and equation (5.9) for all points.

Equation (5.7) is replaced by

$$\begin{aligned}(6.3)\qquad &\frac{1}{(\Delta x)^2}\left[-\rho^h_{i+\frac{1}{2},j}\left(\chi^h_{i+1,j} - \chi^h_{i,j}\right) + \rho^h_{i-\frac{1}{2},j}\left(\chi^h_{i,j} - \chi^h_{i-1,j}\right)\right] \\ &\quad + \frac{1}{(\Delta y)^2}\left[-\rho^h_{i,j+\frac{1}{2}}\left(\chi^h_{i,j+1} - \chi^h_{i,j}\right) + \rho^h_{i,j-\frac{1}{2}}\left(\chi^h_{i,j} - \chi^h_{i,j-1}\right)\right] = -\omega^h_{i,j}\end{aligned}$$

for interior points, that is, for

$$\begin{gathered}1 \leqq i \leqq I - 1 \\ 1 \leqq j \leqq J - 1\end{gathered}$$

where

$$(6.4)\qquad \begin{aligned}2\rho^h_{i\pm\frac{1}{2},j} &= \rho^h_{i\pm1,j} + \rho^h_{i,j} \\ 2\rho^h_{i,j\pm\frac{1}{2}} &= \rho^h_{i,j\pm1} + \rho^h_{i,j}\end{aligned}$$

and $\omega^h_{i,j}$ is formed from $\psi^h_{i,j}$ and $\rho^h_{i,j}$ as indicated in equation (A.14) of Appendix I.

Equation (5.6) is replaced by the equation

$$(6.5)\qquad \psi^{h+1}_{i,j} = \psi^{h-1}_{i,j} - 2\Delta t\chi^h_{i,j}$$

for

$$h > 0$$

and by

$$(6.6)\qquad \psi^1 = \psi^0_{i,j} - \Delta t\chi^0_{i,j}$$

when

$$h = 0$$

Equation (5.9) is replaced by the equation

$$\begin{aligned}(6.7)\qquad \rho^{h+1}_{i,j} = \rho^h_{i,j} &+ \left(\frac{\Delta t}{\Delta x}\right)\left\{\begin{matrix}(\rho^h_{i,j} - \rho^h_{i-1,j}) & \text{if} & (\psi_y)^{h+\frac{1}{2}}_{i,j} < 0 \\ (\rho^h_{i+1,j} - \rho^h_{i,j}) & \text{if} & (\psi_y)^{h+\frac{1}{2}}_{i,j} \geqq 0\end{matrix}\right\}(\psi_y)^{h+\frac{1}{2}}_{i,j} \\ &- \left(\frac{\Delta t}{\Delta y}\right)\left\{\begin{matrix}(\rho^h_{i,j} - \rho^h_{i,j-1}) & \text{if} & (\psi_x)^{h+\frac{1}{2}}_{i,j} \geqq 0 \\ (\rho^h_{i,j+1} - \rho^h_{i,j}) & \text{if} & (\psi_x)^{h+\frac{1}{2}}_{i,j} < 0\end{matrix}\right\}(\psi_x)^{h+\frac{1}{2}}_{i,j}\end{aligned}$$

for interior points, where

$$2(\psi_x)^{h+\frac{1}{2}}_{i,j} = (\psi_x)^{h+1}_{i,j} + (\psi_x)^h_{i,j}$$

and a similar equation defines $(\psi_y)^{h+\frac{1}{2}}_{i,j}$.

On a boundary such as $j = J$, $(\psi_x)^h_{i,j} = 0$ from the boundary conditions, and if initially $\rho^0_{i,J} =$ constant, it will follow from equation (6.7) that $\rho^h_{i,J} = \rho^0_{i,J}$.

That is, the density discontinuity will not be able to reach the boundary $j = J$. For this reason equation (6.7) does not seem to be a suitable one for determining the time behavior of the density at the boundaries. It was replaced by

$$\begin{aligned}\rho_{i,j}^{h+1} = \rho_{i,j}^{h} &+ \frac{\Delta t}{\Delta x}\left\{\begin{matrix} \rho_{i,j}^{h}\,(\psi_y)_{i,j}^{h+\frac{1}{2}} - \rho_{i-1,j}^{h}\,(\psi_y)_{i-1,j}^{h+\frac{1}{2}} & \text{if} \quad (\psi_y)_{i,j}^{h+\frac{1}{2}} < 0 \\ \rho_{i+1}^{h}\,(\psi_y)_{i+1,j}^{h+\frac{1}{2}} - \rho_{i,j}^{h}\,(\psi_y)_{i,j}^{h+\frac{1}{2}} & \text{if} \quad (\psi_y)_{i,j}^{h+\frac{1}{2}} \geqq 0 \end{matrix}\right\} \\ &- \frac{\Delta t}{\Delta y}\left\{\begin{matrix} \rho_{i,j}^{h}\,(\psi_x)_{i,j}^{h+\frac{1}{2}} - \rho_{i,j-1}^{h}\,(\psi_x)_{i,j-1}^{h+\frac{1}{2}} & \text{if} \quad (\psi_x)_{i,j}^{h+\frac{1}{2}} \geqq 0 \\ \rho_{i,j+1}^{h}\,(\psi_x)_{i,j+1}^{h+\frac{1}{2}} - \rho_{i,j}^{h}\,(\psi_x)_{i,j}^{h+\frac{1}{2}} & \text{if} \quad (\psi_x)_{i,j}^{h+\frac{1}{2}} < 0 \end{matrix}\right\}\end{aligned} \tag{6.8}$$

for boundary points. This equation is a finite difference representation of

$$\rho_t = -(\rho u)_x - (\rho v)_y$$

just as equation (6.7) is of equation (5.9).

Equation (6.7) differs from the finite difference form of equation (5.9) proposed by von Neumann, namely

$$\rho_{i,j}^{h+1} = \rho_{i,j}^{h} + \Delta t(\rho_t)_{i,j}^{h} + \frac{(\Delta t)^2}{2}(\rho_{tt})_{i,j}^{h} \tag{6.9}$$

where $(\rho_t)_{i,j}^{h}$ and $(\rho_{tt})_{i,j}^{h}$ are evaluated from the values of $\rho_{i,j}^{h}$, $\psi_{i,j}^{h}$ and $\chi_{i,j}^{h}$ by substituting into the centered finite difference representation of equation (5.9) and the equation obtained by differentiating this equation with respect to t and substituting for ρ_t from (5.9) and χ for $-\psi_t$.

In the early calculations the finite difference form of equation (6.9) was used. However, it was found that near a discontinuity in the density ρ the values of ρ increased on the high side of the discontinuity and decreased on the low side, thus steadily increasing the size of the discontinuity. This unstable behavior did not occur when equation (6.7) was used.

7. The solution of Equation (6.3). This equation is of the form of a set of linear equations which may be written as

$$A\chi = \omega \tag{7.1}$$

where χ is an unknown $M[= (I - 1)(J - 1)]$ dimensional vector, ω is a known vector of this many dimensions, and A is a known $M \times M$ matrix.

In Appendix II von Neumann discusses iteration schemes for solving these equations. He concludes that if A is a positive (or negative) definite matrix [the matrix A of equation (6.3) is negative definite] with a largest proper value less than or equal to b and a smallest one greater than or equal to a, then the "best" (in the sense defined in Appendix II) iteration scheme is given by the equation

$$\eta^{k+1} = 2b_{k+1}\left[\eta^{k} - \eta^{k-1} + \frac{2}{a+b}(\omega - A\eta^{k})\right] + \eta^{k-1} \tag{7.2}$$

where

$$\begin{aligned} b_1 &= 1 \\ b_{k+1} &= \frac{1}{2 - (1-\epsilon)^2 b_k} \end{aligned} \tag{7.3}$$

and

$$\epsilon = \frac{2a}{a+b} \tag{7.4}$$

In order to apply this scheme, bounds for the lowest and highest proper values, the numbers a and b, must be determined.

Using the equations given by von Neumann in Section 15 of Appendix II, we may set

$$\begin{aligned} a &= 4\bar{a}\left(\frac{1}{(\Delta x)^2}\sin^2\frac{\pi}{2I} + \frac{1}{(\Delta y)^2}\sin^2\frac{\pi}{2J}\right) \\ b &= 4\bar{b}\left(\frac{1}{(\Delta x)^2}\cos^2\frac{\pi}{2I} + \frac{1}{(\Delta y)^2}\cos^2\frac{\pi}{2J}\right) \end{aligned} \tag{7.5}$$

where $\bar{a}$ and $\bar{b}$ are lower and upper bounds, respectively, of the density. That is

$$0 < \bar{a} \leqq \rho_{i,j} \leqq \bar{b} \tag{7.6}$$

for all i and j.

8. The Flow Diagram. A condensed flow diagram is reproduced as Fig. 2. Before operation, initial values of $\psi_{i,j}$ and $\rho_{i,j}$ at time $h = 0$ are stored. The values of $\rho_{i,j}$ were prepared as follows: If i, j labels a mesh point which does not lie on the density discontinuity, the value $\rho_{i,j} = \rho(x_i, y_j)$. If x_i, y_j lies on the density discontinuity, we define

$$\rho_{i,j} = \tfrac{1}{2}[\rho(x_i, y_{j+1}) + \rho(x_i, y_{j-1})]$$

$\psi_{i,j}$ was taken to be zero initially. When the routine is started, part A is traversed, which sets $h = 0$ and optionally prints out the initial values of ψ and ρ in a format uniform with subsequent results.

Part B computes the values of $\omega^h_{i,j}$, the right side of equation (5.1), for all values of i, j corresponding to interior points, as needed in equation (7.1). The box with the sole notation i, j indicates an induction loop repeatedly using the program of the box to right of it for all appropriate values of i, j. All derivatives in the formula for $\omega^h_{i,j}$ are computed by taking the difference of the values of the function at lattice points on each side, for example

$$(\psi_x)_{i,j} = \frac{1}{2\Delta x}(\psi_{i+1,j} - \psi_{i-1,j})$$

except in certain cases near the boundary where one of these quantities does not exist, and then a one-sided derivative, e.g.,

$$\frac{1}{\Delta x}(\psi_{i+1,j} - \psi_{i,j}) \qquad \text{for} \quad i = 0, j = 0, 1, 2, \cdots, J$$

is used.

Part C computes η^0, the first term in the sequence of vectors η^k to be constructed converging to χ. If $h = 0$, then η^0 is made zero, which is a reasonable estimate in the case where the liquid starts moving from rest. After the motion has proceeded

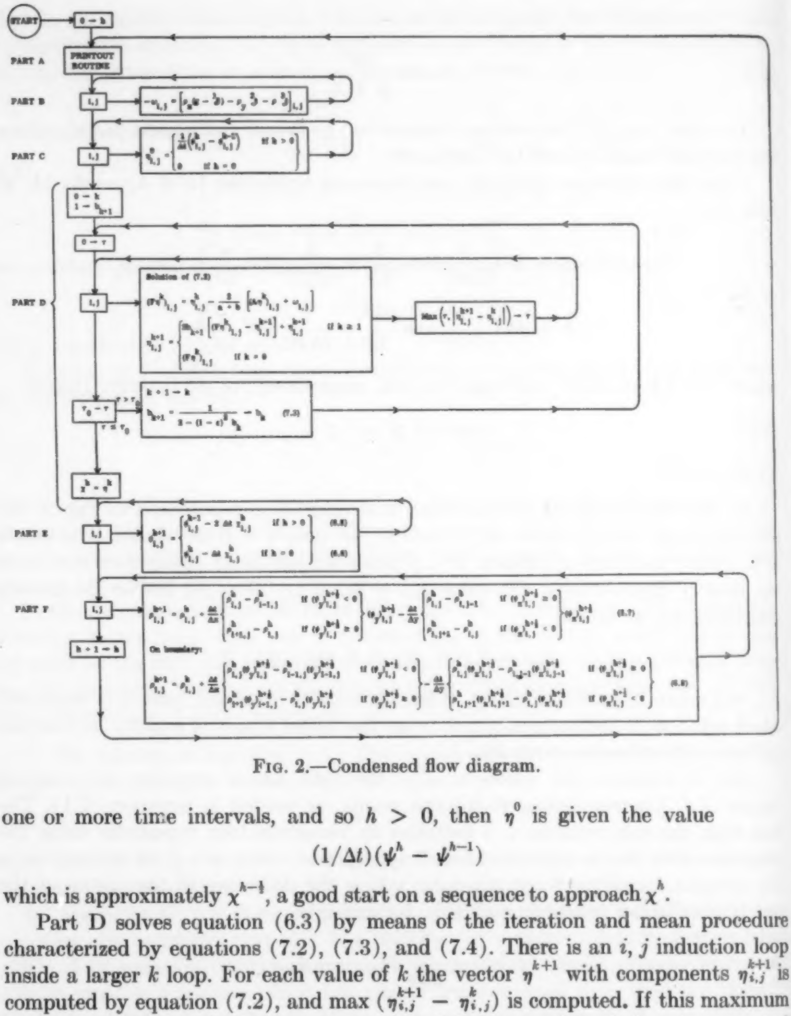

FIG. 2.—Condensed flow diagram.

one or more time intervals, and so $h > 0$, then η^0 is given the value

$$(1/\Delta t)(\psi^h - \psi^{h-1})$$

which is approximately $\chi^{h-\frac{1}{2}}$, a good start on a sequence to approach χ^h.

Part D solves equation (6.3) by means of the iteration and mean procedure characterized by equations (7.2), (7.3), and (7.4). There is an i, j induction loop inside a larger k loop. For each value of k the vector η^{k+1} with components $\eta_{i,j}^{k+1}$ is computed by equation (7.2), and max $(\eta_{i,j}^{k+1} - \eta_{i,j}^{k})$ is computed. If this maximum is above a predetermined constant τ_0, then k is increased by 1, b_{k+1} is computed from equation (7.3), and the i, j induction loop is entered again to compute the next term in the sequence of η^k values. When a vector η^k is obtained which is sufficiently close to η^{k-1}, then it is considered to be χ^h, and the control passes to part E.

Part E computes the new values ψ^{h+1} by equations (6.5) and (6.6), and Part F computes the new values ρ^{h+1} by equations (6.7) and (6.8). The time index h is then increased by 1, the results are optionally printed, and the control passes again to Part B.

The finite difference equations used in the numerical computation according to the flow diagram are:

PART A. Part A has no formulae.

PART B. Part B is an induction loop for computing all interior points of $\omega^h_{i,j}$ as a function of $\psi^h_{i,j}$ and $\rho^h_{i,j}$. The formulae used are

$$2(\Delta x)(\psi_x)^h_{i,j} = \psi^h_{i+1,j} - \psi^h_{i-1,j} \tag{B.1}$$

$$2(\Delta y)(\psi_y)^h_{i,j} = \psi^h_{i,j+1} - \psi^h_{i,j-1} \tag{B.2}$$

$$4(\Delta x)(\Delta y)(\psi_{xy})^h_{i,j} = \psi^h_{i+1,j+1} - \psi^h_{i-1,j+1} - \psi^h_{i+1,j-1} + \psi^h_{i-1,j-1} \tag{B.3}$$

The next three formulae are computed as a five-cycle induction loop for $(i', j') =$ $(i, j), (i + 1, j), (i - 1, j), (i, j + 1), (i, j - 1)$

$$(\Delta x)^2(\psi_{xx})^h_{i',j'} = \psi^h_{i'+1,j'} - 2\psi^h_{i',j'} + \psi^h_{i'-1,j'} \tag{B.4}$$

$$(\Delta y)^2(\psi_{yy})^h_{i',j'} = \psi_{i',j'+1} - 2\psi^h_{i',j'} + \psi^h_{i',j'-1} \tag{B.5}$$

$$(\Delta x)^2\lambda^h_{i',j'} = (\Delta x)^2(\psi_{xx})^h_{i',j'} + a(\Delta y)^2(\psi_{yy})^h_{i',j'} \tag{B.6}$$

where $a = (\Delta x/\Delta y)^2$.

$$2(\Delta x)^3(\lambda_x)^h_{i,j} = (\Delta x)^2\lambda^h_{i+1,j} - (\Delta x)^2\lambda^h_{i-1,j} \quad \text{if} \quad i \neq 1,\ I - 1 \tag{B.7a}$$

$$(\Delta x)^3(\lambda_x)^h_{i,j} = (\Delta x)^2\lambda^h_{i+1,j} - (\Delta x)^2\lambda^h_{i,j} \quad \text{if} \quad i = 1 \tag{B.7b}$$

$$(\Delta x)^3(\lambda_x)^h_{i,j} = (\Delta x)^2\lambda^h_{i,j} - (\Delta x)^2\lambda^h_{i-1,j} \quad \text{if} \quad \mathrm{i} = I - 1 \tag{B.7c}$$

$$2(\Delta x)^2(\Delta y)(\lambda_y)^h_{i,j} = (\Delta x)^2\lambda^h_{i,j+1} - (\Delta x)^2\lambda^h_{i,j-1} \quad \text{if} \quad j \neq 1,\ J - 1 \tag{B.8a}$$

$$(\Delta x)^2(\Delta y)(\lambda_y)^h_{i,j} = (\Delta x)^2\lambda^h_{i,j+1} - (\Delta x)^2\lambda^h_{i,j} \quad \text{if} \quad j = 1 \tag{B.8b}$$

$$(\Delta x)^2(\Delta y)(\lambda_y)^h_{i,j} = (\Delta x)^2\lambda^h_{i,j} - (\Delta x)^2\lambda^h_{i,j-1} \quad \text{if} \quad j = J - 1 \tag{B.8c}$$

$$2(\Delta x)(\rho_x)^h_{i,j} = \rho^h_{i+1,j} - \rho^h_{i-1,j} \tag{B.9}$$

$$2(\Delta y)(\rho_y)^h_{i,j} = \rho^h_{i,j+1} - \rho^h_{i,j-1} \tag{B.10}$$

$$\begin{aligned} 8(\Delta x)^2(\Delta y)\,{}^1I^h_{i,j} = {} & [2(\Delta x)(\psi_x)^h_{i,j}][4(\Delta x)(\Delta y)(\psi_{xy})^h_{i,j}] \\ & - 4[2(\Delta y)(\psi_y)^h_{i,j}][(\Delta x)^2(\psi_{xx})^h_{i,j}] \end{aligned} \tag{B.11}$$

$$\begin{aligned} 8(\Delta x)(\Delta y)^2\,{}^2I^h_{i,j} = {} & 4[2(\Delta x)(\psi_x)^h_{i,j}][(\Delta y)^2(\psi_{yy})^h_{i,j}] \\ & - [2(\Delta y)(\psi_y)^h_{i,j}][4(\Delta x)(\Delta y)(\psi_{xy})^h_{i,j}] \end{aligned} \tag{B.12}$$

$$\begin{aligned} 4(\Delta x)^3(\Delta y)\,{}^3I^h_{i,j} = {} & [2(\Delta x)(\psi_x)^h_{i,j}][2(\Delta x)^2(\Delta y)(\lambda_y)^h_{i,j}] \\ & - [2(\Delta y)(\psi_y)^h_{i,j}][2(\Delta x)^3(\lambda_x)^h_{i,j}] \end{aligned} \tag{B.13}$$

$$\begin{aligned} 16(\Delta x)^3(\Delta y)\omega^h_{i,j} = {} & [2(\Delta x)(\rho_x)^h_{i,j}][8(\Delta x)^2(\Delta y)\,{}^1I^h_{i,j} - 8(\Delta x)^2(\Delta y)g] \\ & + a[2(\Delta y)(\rho_y)^h_{i,j}][8(\Delta x)(\Delta y)^2\,{}^2I^h_{i,j}] + [4\rho^h_{i,j}][4(\Delta x)^3(\Delta y)\,{}^3I^h_{i,j}] \end{aligned} \tag{B.14}$$

where

$$a = \left(\frac{\Delta x}{\Delta y}\right)^2$$

Part C. Part C is an induction loop, with respect to i, j of all lattice points to compute the first trial value, $\eta^0_{i,j}$, at cycle h for use in the iteration process. The formula is

$$\Delta x \Delta y\, \eta^0_{i,j} = \left\{ \begin{array}{ll} \dfrac{(\Delta x)(\Delta y)}{\Delta t}(\psi^{h-1}_{i,j} - \psi^h_{i,j}) & \text{if} \quad h > 0 \\ 0 & \text{if} \quad h = 0 \end{array} \right\} \tag{C.1}$$

Part D. Part D is an induction loop with respect to k, with a smaller i, j induction loop for each k. This is the solution of the difference equation by the iterative and mean method. The actual formulae are

$$2\rho^h_{i+\frac{1}{2},j} = \rho^h_{i,j} + \rho^h_{i+1,j} \tag{D.1}$$

$$2\rho^h_{i-\frac{1}{2},j} = \rho^h_{i,j} + \rho^h_{i-1,j} \tag{D.2}$$

$$2a\rho^h_{i,j+\frac{1}{2}} = a(\rho^h_{i,j} + \rho^h_{i,j+1}) \tag{D.3}$$

As before

$$a = \left(\frac{\Delta x}{\Delta y}\right)^2$$

$$2a\rho^h_{i,j-\frac{1}{2}} = a(\rho^h_{i,j} + \rho^h_{i,j-1}) \tag{D.4}$$

$$\sigma^h_{i,j} = 2\rho^h_{i+\frac{1}{2},j} + 2\rho^h_{i-\frac{1}{2},j} + 2a\rho^h_{i,j+\frac{1}{2}} + 2a\rho^h_{i,j-\frac{1}{2}} \tag{D.5}$$

$$\begin{aligned} -2(\Delta x)^3(\Delta y)(A\eta^k)_{i,j} = {} & [2\rho^h_{i+\frac{1}{2},j}][(\Delta x)(\Delta y)\eta^k_{i+1,j}] \\ & + [2\rho^h_{i-\frac{1}{2},j}][(\Delta x)(\Delta y)\eta^k_{i-1,j}] + [2a\rho^h_{i,j+\frac{1}{2}}][(\Delta x)(\Delta y)\eta^k_{i,j+1}] \\ & + [2a\rho^h_{i,j-\frac{1}{2}}][(\Delta x)(\Delta y)\eta^k_{i,j-1}] - \sigma_{i,j}[(\Delta x)(\Delta y)\eta^k_{i,j}] \end{aligned} \tag{D.6}$$

$$\begin{aligned} (\Delta x)(\Delta y)(F\eta^k)_{i,j} = {} & [(\Delta x)(\Delta y)\eta^k_{i,j}] + \frac{1}{2}\left[\frac{a}{(\Delta x)^2}\right]([-2(\Delta x)^3(\Delta y)(A\eta^k)_{i,j}] \\ & - \frac{1}{8}[16(\Delta x)^3(\Delta y)\omega^h_{i,j}]) \end{aligned} \tag{D.7}$$

where $d \equiv \dfrac{\epsilon}{a}$, cf. Eq. (7.4).

$$(\Delta x)(\Delta y)\eta^{k+1}_{i,j} = \left\{ \begin{array}{l} 2b_{k+1}\{[(\Delta x)(\Delta y)(F\eta^k)_{i,j}] - [(\Delta x)(\Delta y)\eta^{k-1}_{i,j}]\} \\ \qquad\qquad + [(\Delta x)(\Delta y)\eta^{k-1}_{i,j}] \quad \text{if} \quad k = 0 \\ (\Delta x)(\Delta y)(F\eta^k)_{i,j} \quad \text{if} \quad k = 0 \end{array} \right\} \tag{D.8}$$

$$b_1 = 1 \qquad \text{for} \quad k = 0 \tag{D.9}$$

$$b_{k+1} = \frac{1}{2 - (1-\epsilon)^2 b_k} \qquad \text{for} \quad k > 0 \tag{D.10}$$

Part E. Part E is an induction loop with respect to i, j and is used to compute $\psi^{h+1}_{i,j}$ for all points as a function of $\psi^{h-1}_{i,j}$ and $\chi^h_{i,j}$

$$\psi^{h+1}_{i,j} = \psi^{h-1}_{i,j} - \frac{2\Delta t}{(\Delta x)(\Delta y)}[(\Delta x)(\Delta y)\chi^h_{i,j}] \qquad \text{if} \quad h > 0 \tag{E.1a}$$

$$\psi_{i,j}^{h+1} = \psi_{i,j}^{h} - \frac{\Delta t}{(\Delta x)(\Delta y)}\,[(\Delta x)(\Delta y)\chi_{i,j}^{h}] \qquad \text{if} \quad h = 0 \tag{E.1b}$$

PART F. Part F is an induction loop with respect to i, j and is used to compute $\rho_{i,j}^{h+1}$ as a function of $\psi_{i,j}^{h}$, $\psi_{i,j}^{h+1}$, and $\rho_{i,j}^{h}$.

$$\begin{aligned} \rho_{i,j}^{h+1} = \rho_{i,j}^{h} &+ \frac{\Delta t}{\Delta x}\left\{\begin{matrix} (\rho_{i,j}^{h} - \rho_{i-1,j}^{h}) & \text{if} & (\psi_y)_{i,j}^{h+\frac{1}{2}} < 0 \\ (\rho_{i+1,j}^{h} - \rho_{i,j}^{h}) & \text{if} & (\psi_y)_{i,j}^{h+\frac{1}{2}} \geqq 0 \end{matrix}\right\}(\psi_y)_{i,j}^{h+\frac{1}{2}} \\ &- \frac{\Delta t}{\Delta y}\left\{\begin{matrix} (\rho_{i,j}^{h} - \rho_{i,j-1}^{h}) & \text{if} & (\psi_x)_{i,j}^{h+\frac{1}{2}} \geqq 0 \\ (\rho_{i,j+1}^{h} - \rho_{i,j}^{h}) & \text{if} & (\psi_x)_{i,j}^{h+\frac{1}{2}} < 0 \end{matrix}\right\}(\psi_x)_{i,j}^{h+\frac{1}{2}} \end{aligned} \tag{F.1a}$$

$$\begin{aligned} \text{for} \quad & i = 0, 1, \cdots, I \\ & j = 1, 2, \cdots, J-1 \end{aligned}$$

$$\begin{aligned} \rho_{i,j}^{h+1} = \rho_{i,j}^{h} &+ \frac{\Delta t}{\Delta x}\left\{\begin{matrix} \rho_{i,j}^{h}(\psi_y)_{i,j}^{h+\frac{1}{2}} - \rho_{i-1,j}^{h}(\psi_y)_{i-1,j}^{h+\frac{1}{2}} & \text{if} & (\psi_y)_{i,j}^{h+\frac{1}{2}} < 0 \\ \rho_{i+1,j}^{h}(\psi_y)_{i+1,j}^{h+\frac{1}{2}} - \rho_{i,j}^{h}(\psi_y)_{i,j}^{h+\frac{1}{2}} & \text{if} & (\psi_y)_{i,j}^{h+\frac{1}{2}} \geqq 0 \end{matrix}\right\} \\ &- \frac{\Delta t}{\Delta y}\left\{\begin{matrix} \rho_{i,j}^{h}(\psi_x)_{i,j}^{h+\frac{1}{2}} - \rho_{i,j-1}^{h}(\psi_x)_{i,j-1}^{h+\frac{1}{2}} & \text{if} & (\psi_x)_{i,j}^{h+\frac{1}{2}} \geqq 0 \\ \rho_{i,j+1}^{h}(\psi_x)_{i\;j+1}^{h+\frac{1}{2}} - \rho_{i,j}^{h}(\psi_x)_{i,j}^{h+\frac{1}{2}} & \text{if} & (\psi_x)_{i,j}^{h+\frac{1}{2}} < 0 \end{matrix}\right\} \end{aligned} \tag{F.1b}$$

$$\begin{aligned} \text{for} \quad & i = 1, 2, \cdots, I-1 \\ & j = 0, J \end{aligned}$$

9. Stability and the Choice of Δt. The behavior of the solutions of equation (6.7) with regard to stability is similar to that of the corresponding equation in one dimension with a constant velocity of propagation which will be taken to be positive. Then the equation of conservation of mass becomes

$$\rho_t = -u\rho_x$$

which has the finite difference representation

$$\rho_j^{h+1} = \rho_j^{h} - \frac{\Delta t}{\Delta x}\,u(\rho_j^{h} - \rho_{j-1}^{h})$$

or

$$\rho_j^{h+1} = (1-\alpha)\rho_j^{h} + \alpha\rho_{j-1}^{h} \tag{9.1}$$

where

$$\alpha = \frac{u\Delta t}{\Delta x} \tag{9.2}$$

Equation (9.1) has solutions of the form

$$\rho_j^{h} = \exp\left[\frac{i\pi\eta}{J}\,(j\Delta x - \beta h\Delta t)\right] \tag{9.3}$$

where

$$\exp\left[-\frac{i\pi\eta}{J}\,\beta\Delta t\right] = 1 + \alpha\exp\left[-\frac{i\pi\eta}{J}\,\Delta x - 1\right]$$

and hence

$$\left|\exp\left(-\frac{i\pi\eta}{J}\beta\Delta t\right)\right|^2 = 1 - 4\alpha(1 - \alpha)\cos^2\pi\frac{\eta\Delta x}{2J} \tag{9.4}$$

Thus β will be real, and the equation will be stable if and only if

$$\alpha \leqq 1$$

That is, if

$$\Delta t \leqq \frac{\Delta x}{u}$$

The value of Δt in the two-dimensional calculations was chosen so that

$$|\psi_y| \leqq \frac{\Delta x}{\Delta t}$$

and

$$|\psi_x| \leqq \frac{\Delta y}{\Delta t}$$

The units used were such that a change in Δt could be accomplished by changing the value of the constant representing g the acceleration of gravity, and scaling ψ.

It follows from equation (9.4) that up to second-order terms in Δx

$$\beta = u\left[1 - i\left(1 - u\frac{\Delta t}{\Delta x}\right)\frac{\pi\eta}{2J}\Delta x\right] \tag{9.5}$$

Hence the elementary solutions of equation (9.1) of the form of (9.3) may be written as

$$\rho_j = \exp\left[\frac{i\pi\eta}{J}(j\Delta x - h\Delta tu)\right]\exp\left[-j\left(\frac{\pi\eta}{J}\right)^2\Delta x^2 h\alpha(1 - \alpha)\right] \tag{9.6}$$

The first factor of this expression may be written as

$$\exp\left[\frac{i\pi\eta}{J}(x - ut)\right]$$

with $x = j\Delta x$ and $t = h\Delta t$. This is a solution of the differential equation. Thus the second factor on the right-hand side of equation (9.6) shows how each of these elementary solutions of the differential equation is distorted when that equation is replaced by the finite difference equation (9.1).

Von Neumann in a personal communication to S. Ulam (cf. Appendix III of reference 1) has used the term "pseudo-diffusion" for the distortion associated with an initial distribution of density

$$\rho_0(x) = \begin{Bmatrix} 1 & \text{for} & x \geqq 0 \\ 0 & \text{for} & x < 0 \end{Bmatrix}$$

He has shown that if this function is taken as an initial condition for the difference equation (9.1), then the 0 values of ρ advance and the 1 values of ρ recede to the

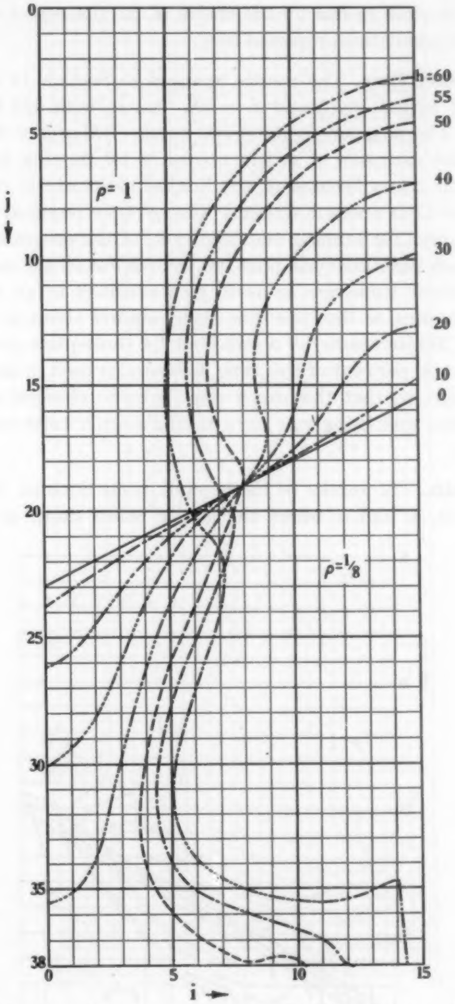

FIG. 3.—Gravity flow of two incompressible inviscid fluids at various cycles h.

right with a velocity

$$\frac{\delta x}{\delta t} = u$$

and at the same time a pseudo-diffusion-mixing region forms around the interface of advance whose width is essentially measured by

$$\delta x \sim \sqrt{(1 - \alpha)x \cdot \Delta x}$$

Since α was made close to one by the choice of Δt, the region of pseudo-diffusion was small for the calculations reported here.

10. Computation Time. The results reported in Section 11 of this paper were run on Maniac I with $I = 15$ and $J = 38$, that is, with 624 lattice points (518 interior points). The program required 300 words (600 orders of code exclusive of print routines and exclusive of orders necessary for moving information to and from the magnetic drum because of the limited electrostatic storage capacity of Maniac I). About 3750 words of dynamic storage were required.

The time required for running one time cycle of the program on Maniac I was 18 seconds for each iteration cycle plus 100 seconds for all the rest of the program. The iteration process converges so as to give accuracy in an additional decimal place every 10 minutes, so that one time cycle requires about an hour for six-place accuracy (about 200 iterations) or a half hour for three-place accuracy (about 100 iterations). About 40 per cent of this time, however, is used in transfers to and from the magnetic drum, so that this much time is to be charged to the fact that a 4000-word problem was being run on a machine with 1000-word random-access memory capacity.

11. The Results. The results of the computations done on Maniac I are summarized in Figs. 3, 4, and 5, where the lattice points occur at integer values of

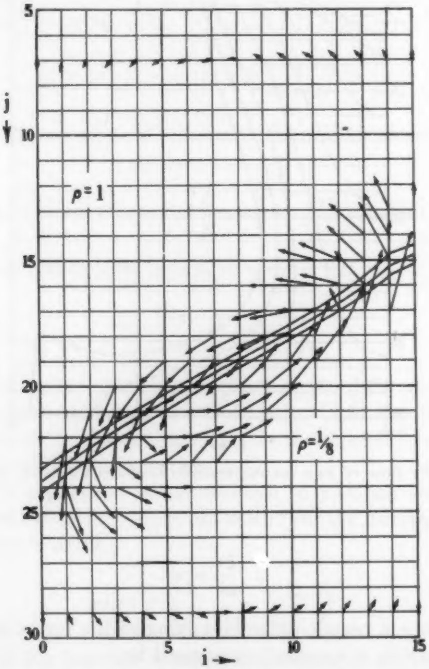

Fig. 4.—Velocity field at $h = 10$.

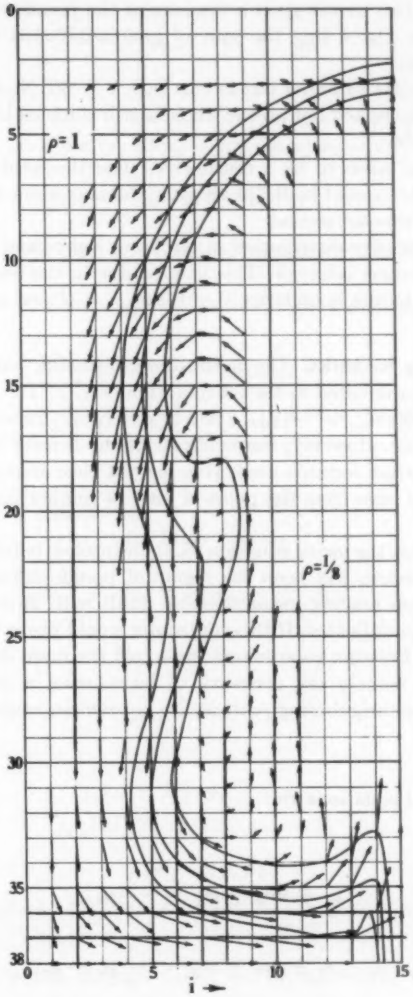

FIG. 5.—Velocity field at $h = 60$.

the abscissae (i) and ordinates (j). In the first of these the locus of $\rho = \frac{1}{2}(\rho_{\text{upper}} + \rho_{\text{lower}}) = 0.5625$ is shown for various values of h, where $h\Delta t$ is the time elapsed since the start of the calculations. Curves are shown for $h = 0, 10, 20, 30, 40, 50, 55,$ and 60.

Figure 4 shows the velocity field at $h = 10$, as well as the loci $\rho = 0.5625$, and $\rho = 0.3875$, and $\rho = 0.7375$. The latter two curves are the loci at which 30 and 70 per cent, respectively, of the initial difference in density are achieved. The band covered by these two curves gives a measure of the pseudo-diffusion phenomenon. It is evident from Fig. 4 that the zone of pseudo-diffusion is less than one mesh length in thickness.

Figure 5 is similar to Fig. 4 but in this case $h = 60$. Now the pseudo-diffusion phenomenon has increased but on the whole is still confined to a region of the order of two mesh lengths.

If $\Delta x = \Delta y$ is taken to be 1 centimeter, then the total time covered by the calculations ($60\ \Delta t$) would be 0.339 second. The maximum speed attained by the fluid is 0.0184 centimeter/second.

In Fig. 5 the density distribution in the lower right-hand corner seems to have a somewhat anomalous behavior. This may be due to the fact that equation (6.8) was only applied to the boundaries $j = 0$ and $j = J$ and not to the boundaries $i = 0$ and $i = I$.

12. Concluding Remarks. The most time-consuming part of the calculations performed was that devoted to the computation of $\chi^{h}_{i,j}$. The subsequent computation of the velocities $u^{h}_{i,j} = -(\psi_y)^{h}_{i,j}$ and $v^{h}_{i,j} = (\psi_x)^{h}_{i,j}$, and the density $\rho^{h}_{i,j}$ took relatively little time. However, the behavior of the density was most sensitive to the type of integration formula used. It is not yet clear that the formulas actually used were the best ones from the point of view of minimizing the zone of pseudo-diffusion.

It is expected that the use of equation (6.8) for interior points as well as boundary points or other devices will keep the region of pseudo-diffusion small enough so that calculations on moving incompressible fluids with moving interfaces can be made in Eulerian coordinates. If this conjecture would prove to be correct it would be possible to use Eulerian coordinates and avoid the main difficulty of working in Lagrangian ones; namely the necessity of introducing a new Lagrangian mesh periodically because neighboring particles do not remain neighboring particles.

APPENDIX I*

The differential equations are:

Interior:

$$u_t + uu_x + vu_y = -\frac{1}{\rho} p_x, \tag{1}$$

$$v_t + uv_x + vv_y = -\frac{1}{\rho} p_y + g, \tag{2}$$

$$\rho_t + u\rho_x + v\rho_y = 0, \tag{3}$$

$$u_x + v_y = 0. \tag{4}$$

Boundary:

$$\cos(x, n)\, u + \cos(y, n)\, v = 0. \tag{5}$$

* Although the material in Appendix I and Appendix II was left by von Neumann in the form of handwritten notes and not in a form intended for wide distribution, the value of its content is thought to justify its inclusion in this paper.

From (4):

(6) $$u = -\psi_y, \qquad v = \psi_x.$$

(6) replaces (4).

Now (5) becomes:

$$-\cos(x, n)\, \psi_y + \cos(y, n)\, \psi_x = 0,$$

i.e.,

$$\psi_x : \psi_y = \cos(x, n) : \cos(y, n),$$

i.e.,

$$\psi_x, \psi_y \quad \text{is purely normal},$$

i.e.,

$$\psi_{\tan} = 0 \quad \text{(the tangential derivative of } \psi \text{ vanishes)}.$$

This means that

$$\psi = C \ (= \text{constant})$$

along the boundary. Now replacing ψ by $\psi - C$ does not interfere with (6) (ψ's defining relation). Hence $C = 0$ may be assumed, i.e.:

(7) $$\psi = 0$$

(on the boundary). (7) replaces (5).

Now only (1), (2), (3) are left. These become:

$$-\psi_{yt} + \psi_y\, \psi_{xy} - \psi_x\, \psi_{yy} = -\frac{1}{\rho}\, p_x,$$

$$\psi_{xt} - \psi_y\, \psi_{xx} + \psi_x\, \psi_{xy} = -\frac{1}{\rho}\, p_y + g,$$

$$\rho_t - \psi_y\, \rho_x + \psi_x\, \rho_y = 0,$$

i.e.,

(8) $$\rho[\psi_{ty} + I(\psi, \psi_y)] = p_x,$$

(9) $$\rho[\psi_{tx} + I(\psi, \psi_x) - g] = -p_y,$$

(10) $$\rho_t = -I(\psi, \rho).$$

(8), (9), (10) replace (1), (2), (3). p can be eliminated between (8), (9), by forming $(8)_y + (9)_x$. This gives

$$\{\rho[\psi_{ty} + I(\psi, \psi_y)]\}_y + \{\rho[\psi_{tx} + I(\psi, \psi_x) - g]\}_x = 0,$$

i.e.,

(11) $$(\rho\psi_{tx})_x + (\rho\psi_{ty})_y = -\{-g\rho_x + [\rho I(\psi, \psi_x)]_x + [\rho I(\psi, \psi_y)]_y\}.$$

(11) replaces (8), (9) (with p eliminated).

Thus the entire system now consists of (10), (11) (interior) and (7) (boundary), while (6) is merely a definition.

Rewriting (10), (11) and (6):

Interior:

(12) $$L\psi_t = -\{[\rho(-g + I[\psi, \psi_x])]_x + [\rho I(\psi, \psi_y)]_y\},$$

where

$$L\chi \equiv (\rho\chi_x)_x + (\rho\chi_y)_y\,, \tag{13}$$
$$\rho_t = -I(\psi, \rho).$$

Boundary:

$$\psi = 0. \tag{14}$$

This suggests the following procedure: Put

$$\chi = -\psi_t\,. \tag{15}$$

Then:

Let ψ, ρ be known. Then ψ_t, ρ_t are obtained by this procedure: Put

$$\omega = \{\rho[-g + I(\psi, \psi_x)]\}_x + [\rho I(\psi, \psi_y)]_y\,. \tag{16}$$

Determine χ from this elliptic boundary value problem:

$$L\chi = \omega, \tag{17}$$

where

$$L\chi \equiv (\rho\chi_x)_x + (\rho\chi_y)_y\,,$$

and on the boundary

$$\chi = 0. \tag{18}$$

Then:

$$\psi_t = -\chi, \tag{19}$$

$$\rho_t = -I(\psi, \rho). \tag{20}$$

The expression (16) for ω may be rewritten:

$$\omega = \rho_x[-g + I(\psi, \psi_x)] + \rho[I(\psi, \psi_x)]_x + \rho_y I(\psi, \psi_y) + \rho[I(\psi, \psi_y)]_y\,.$$

Now

$$[I(\psi, \psi_x)]_x = I(\psi_x\,, \psi_x) + I(\psi, \psi_{xx}) = I(\psi, \psi_{xx}),$$
$$[I(\psi, \psi_y)]_y = I(\psi_y\,, \psi_y) + I(\psi, \psi_{yy}) = I(\psi, \psi_{yy}).$$

Hence

$$\omega = \rho_x[-g + I(\psi, \psi_x)] + \rho_y I(\psi, \psi_y) + \rho I(\psi, \psi_{xx} + \psi_{yy}). \tag{21}$$

Equation (21) replaces (16); it is more convenient for numerical calculation.

This is a more detailed expression for ω:

$$\lambda = \psi_{xx} + \psi_{yy}\,, \tag{22.1}$$

$${}^1I = \psi_x\psi_{xy} - \psi_y\psi_{xx}\,, \tag{22.2}$$

$${}^2I = \psi_x\psi_{yy} - \psi_y\psi_{xy}\,, \tag{22.3}$$

$${}^3I = \psi_x\lambda_y - \psi_y\lambda_x\,, \tag{22.4}$$

$$\omega = \rho_x(g + {}^1I) + \rho_y{}^2I + \rho{}^3I. \tag{22.5}$$

In addition to this the right hand side of (20) is the negative of

$$^4I = I(\psi, \rho), \tag{23}$$

where in detail

$$^4I = \psi_x\rho_y - \psi_y\rho_x . \tag{24}$$

Thus the relevant equations are these: (17) with (22.1)–(22.5) and (on the boundary) (18), and then (19) and (20) with (24).

Now introduce a finite lattice for x, y, t:

$$x = x_i \equiv i\,\Delta x, \qquad i = 0, 1, \cdots, I - 1, I, \tag{25.1}$$

$$y = y_j \equiv j\,\Delta y, \qquad j = 0, 1, \cdots, J - 1, J, \tag{25.2}$$

$$t = t^h \equiv h\,\Delta t, \qquad h = 0, 1, 2, \cdots . \tag{25.3}$$

Occasionally indices $i \pm \frac{1}{2}$, $j \pm \frac{1}{2}$, $h \pm \frac{1}{2}$ are also used. For any quantity

$$\alpha = \alpha(x, y, t). \tag{26.1}$$

The following convention is used:

$$\alpha^h_{ij} = \alpha(x_i, y_j, t^h). \tag{26.2}$$

(17) and (22.1)–(22.5) are needed for $i = 1, \cdots, I - 1$; $j = 1, \cdots, J - 1$ only. (18) is needed for the other i, j only: $i = 0, I$; $j = 0, 1, \cdots, J - 1, J$ or $i = 0, 1, \cdots, I - 1, I$; $j = 0, J$. (19), (20) are needed for all i, j: $i = 0, 1, \cdots, I - 1, I$; $j = 0, 1, \cdots, J - 1, J$.

The entire system of equations can now be rewritten as follows:

ψ^h_{ij} is defined for $i = 1, \cdots, I - 1$; $j = 1, \cdots, J - 1$.

ρ^h_{ij} is defined for $i = 0, 1, \cdots, I - 1, I$; $j = 0, 1, \cdots, J - 1, J$.

In (A.1)–(A.14):*

$$i = 1, \cdots, I - 1; \qquad j = 1, \cdots, J - 1.$$

$$(\bar{\psi}_x)^h_{ij} = \underset{\sim}{\psi^h_{i+1j}} - \underline{\psi^h_{i-1j}} \tag{A.1}$$

for $i \neq 1, I - 1$;
for $i = 1$ omit term $\underline{\qquad}$;
for $i = I - 1$ omit term $\underset{\sim\sim}{\qquad}$.

$$(\bar{\psi}_y)^h_{ij} = \underset{\sim}{\psi^h_{ij+1}} - \underline{\psi^h_{ij-1}} \tag{A.2}$$

for $j \neq 1, J - 1$;
for $j = 1$ omit term $\underline{\qquad}$;
for $j = J - 1$ omit term $\underset{\sim\sim}{\qquad}$.

$$(\bar{\psi}_{xy})^h_{ij} = \underset{\cdots}{\underset{\sim}{\psi^h_{i+1j+1}}} - \underset{\cdots}{\underline{\psi^h_{i-1j+1}}}\ \underset{---}{\underset{\sim}{\psi^h_{i+1j-1}}} + \underset{---}{\underline{\psi^h_{i-1j-1}}} \tag{A.3}$$

* The bar notation includes the required constant times Δx or Δy combinations.

for $i \neq 1, I-1; j \neq 1, J-1$;
for $i = 1$ omit terms$\underline{\qquad}$;
for $i = I-1$ omit terms$\underset{\sim\sim\sim}{\qquad}$;
for $j = 1$ omit terms$\underset{-\,-\,-}{\qquad}$;
for $j = J-1$ omit terms....

In (A.4)–(A.6):

$$(i', j') = (i, j), \underset{\sim\sim\sim\sim\sim\sim}{(i+1, j)}, \underline{(i-1, j)}, \underset{\cdots\cdots\cdots}{(i, j+1)}, \underset{-\,-\,-\,-\,-\,-}{(i, j-1)}$$

for $i \neq 1, I-1; j \neq 1, J-1$;
for $i = 1$ omit term$\underline{\qquad}$;
for $i = I-1$ omit term$\underset{\sim\sim\sim}{\qquad}$;
for $j = 1$ omit term$\underset{-\,-\,-}{\qquad}$;
for $j = J-1$ omit term....
(Hence $i' = 1, \cdots, I-1; j' = 1, \cdots, J-1$.)

$$(\bar{\psi}_{xx})^h_{i'j'} = \underset{\sim\sim\sim}{\psi^h_{i'+1j'}} - 2\psi^h_{i'j'} + \underline{\psi^h_{i'-1j'}} \tag{A.4}$$

for $i' \neq 1, I-1$;
for $i' = 1$ omit term$\underline{\qquad}$;
for $i' = I-1$ omit term$\underset{\sim\sim\sim}{\qquad}$.

$$(\bar{\psi}_{yy})^h_{i'j'} = \underset{\sim\sim\sim}{\psi^h_{i'j'+1}} - 2\psi^h_{i'j'} + \underline{\psi^h_{i'j'-1}} \tag{A.5}$$

for $j \neq 1, J-1$;
for $j' = 1$ omit term$\underline{\qquad}$;
for $j' = J-1$ omit term$\underset{\sim\sim\sim}{\qquad}$.

$$\bar{\lambda}^h_{i'j'} = (\bar{\psi}_{xx})^h_{i'j'} + a(\bar{\psi}_{yy})^h_{i'j'}, \tag{A.6}$$

where

$$a = \left(\frac{\Delta x}{\Delta y}\right)^2.$$

$$(\bar{\lambda}_x)^h_{ij} = \bar{\lambda}^h_{i+1j} - \bar{\lambda}^h_{i-1j} \tag{A.7}$$

for $i \neq 1, I-1$;
for $i = 1$ replace the index $i - 1$ by 1 and double the entire expression;
for $i = I - 1$ replace the index $i + 1$ by $I - 1$ and double the entire expression.

$$(\bar{\lambda}_y)^h_{ij} = \bar{\lambda}^h_{ij+1} - \bar{\lambda}_{ij-1} \tag{A.8}$$

for $j \neq 1, J-1$;
for $j = 1$ replace the index $j - 1$ by 1 and double the entire expression;
for $j = J - 1$ replace the index $j + 1$ by $J - 1$ and double the entire expression.

$$(\bar{\rho}_x)^h_{ij} = \rho^h_{i+1j} - \rho^h_{i-1j}. \tag{A.9}$$

$$(\bar{\rho}_y)^h_{ij} = \rho^h_{ij+1} - \rho^h_{ij-1}. \tag{A.10}$$

$${}^1\bar{I}^h_{ij} = (\bar{\psi}_x)^h_{ij}(\bar{\psi}_{xy})^h_{ij} - (\bar{\psi}_y)^h_{ij}(\bar{\psi}_{xx})^h_{ij}. \tag{A.11}$$

$${}^2\bar{I}^h_{ij} = (\bar{\psi}_x)^h_{ij}(\bar{\psi}_{yy})^h_{ij} - (\bar{\psi}_y)^h_{ij}(\bar{\psi}_{xy})^h_{ij}. \tag{A.12}$$

(A.13) $$ {}^{3}\bar{I}^{h}_{ij} = (\bar{\psi}_x)^{h}_{ij}(\bar{\lambda}_y)^{h}_{ij} - (\bar{\psi}_y)^{h}_{ij}(\bar{\lambda}_x)^{h}_{ij} . $$

(A.14) $$ \bar{\omega}^{h}_{ij} = (\bar{\rho}_x)^{h}_{ij}\,(-\bar{g} + {}^{1}\bar{I}^{h}_{ij}) + a(\bar{\rho}_y)^{h}_{ij}\,{}^{2}\bar{I}^{h}_{ij} + \bar{\rho}^{h}_{ij}\,{}^{3}\bar{I}^{h}_{ij} , $$

where $\bar{g} = (\Delta x)^2\,\Delta y g$ [for a, cf. (A.6)].

In (B.1)–(B.6):

$$ i = 1,\cdots, I - 1; \qquad j = 1,\cdots, J - 1. $$

This definition of $\bar{\chi}^{h}_{ij}$ [i.e., (B.6)] is, however, implicit.

(B.1) $$ {}^{1}\bar{\sigma}^{h}_{ij} = \bar{\rho}^{h}_{ij} + \bar{\rho}^{h}_{i+1j} . $$

(B.2) $$ {}^{2}\bar{\sigma}^{h}_{ij} = \bar{\rho}^{h}_{ij} + \bar{\rho}^{h}_{i-1j} . $$

(B.3) $$ {}^{3}\bar{\sigma}^{h}_{ij} = a(\bar{\rho}^{h}_{ij} + \bar{\rho}^{h}_{ij+1}). $$

(B.4) $$ {}^{4}\bar{\sigma}^{h}_{ij} = a(\bar{\rho}^{h}_{ij} + \bar{\rho}^{h}_{ij-1}). $$

(B.5) $$ {}^{5}\bar{\sigma}^{h}_{ij} = ({}^{1}\bar{\sigma}^{h}_{ij} + {}^{2}\bar{\sigma}^{h}_{ij}) + a({}^{3}\bar{\sigma}^{h}_{ij} + {}^{4}\bar{\sigma}^{h}_{ij}) \qquad \text{[for } a\text{, cf. (A.6)].} $$

(B.6) $$ \utilde{{}^{1}\bar{\sigma}^{h}_{ij}\,\bar{\chi}^{h}_{i+1j}} + \underline{{}^{2}\bar{\sigma}^{h}_{ij}\,\bar{\chi}^{h}_{i-1j}} + \underset{\cdots\cdots}{{}^{3}\bar{\sigma}^{h}_{ij}\,\bar{\chi}^{h}_{ij+1}} + \underset{------}{{}^{4}\bar{\sigma}^{h}_{ij}\,\bar{\chi}^{h}_{ij-1}} - \underline{\underline{{}^{5}\bar{\sigma}^{h}_{ij}\,\bar{\chi}^{h}_{ij}}} = \bar{\omega}^{h}_{ij} $$

for i $\neq 1, I - 1; j \neq 1, J - 1$;

for i $= 1$ omit the term____;

for i $= I - 1$ omit the term∼∼∼;

for $j = 1$ omit the term- - -;

for $j = J - 1$ omit the term....

The term with the double underscore should be $C\,\bar{\chi}^{h}_{ij}$, the corrected $\bar{\chi}^{h}_{ij}$.

In (C):

$$ i = 1,\cdots, I - 1; \qquad j = 1,\cdots, J - 1. $$

(C) $$ \psi^{h+1}_{ij} = \psi^{h-1}_{ij} - 4b\,\bar{\chi}^{h}_{ij} $$

where

$$ b = \frac{\Delta t}{\Delta x\,\Delta y} . $$

In (D.1–(D.3):

$$ i = 0, 1,\cdots, I - 1, I; \qquad j = 0, 1,\cdots, J - 1, J. $$

As in (C):

(D.1) $$ b = \frac{\Delta t}{\Delta x\,\Delta y} . $$
$$ \alpha = \utilde{\underline{\tfrac{1}{2}b}(\psi^{h}_{ij+1} + \psi^{h+1}_{ij+1}} - \underline{\psi^{h}_{ij-1} - \psi^{h+1}_{ij-1})}, $$

for $i \neq 0, I$ and $j \neq 0, J$;

for $i = 0, I$ omit the entire expression;

for $i \neq 0, I$ and $J = 0$ omit the terms and factor$\underline{\quad\;}$;

for $i \neq 0, I$ and $j = J$ omit the terms and factor$\utilde{\quad\;}$.

$$\beta = \utilde{\underline{\tfrac{1}{2}}b(-\overset{h}{\psi}_{i+1j} - \overset{h+1}{\psi}_{i+1j}} + \underline{\overset{h}{\psi}_{i-1j} + \overset{h+1}{\psi}_{i-1j})} \tag{D.2}$$

for $i \neq 0, I$ and $j \neq 0, J$;

for $j = 0, J$ omit the entire expression;

for $i = 0$ omit the terms and factor$\underline{\quad\;}$;

for $j \neq 0$, J and $i = I$ omit the terms and factor$\utilde{\quad\;}$.

$$\begin{aligned} \overset{h+1}{\rho}_{ij} = {} & \overset{h}{\rho}_{ij}(1 - \underline{\alpha^2} - \utilde{\beta^2}) + \underline{\tfrac{1}{2}\overset{h}{\rho}_{i+1j}(\alpha + \alpha^2)} + \underline{\tfrac{1}{2}\overset{h}{\rho}_{i-1j}(-\alpha + \alpha^2})\\ & + \utilde{\tfrac{1}{2}\overset{h}{\rho}_{ij+1}(\beta + \beta^2)} + \utilde{\tfrac{1}{2}\overset{h}{\rho}_{ij-1}(-\beta + \beta^2})\\ & + \utilde{\underline{\tfrac{1}{4}(\overset{h}{\rho}_{i+1j+1} - \overset{h}{\rho}_{i+1j-1} - \overset{h}{\rho}_{i-1j+1} + \overset{h}{\rho}_{i-1j-1}) - \alpha\beta}} \end{aligned} \tag{D.3}$$

for $i \neq 0, I$ and $j \neq 0, J$;

for $i = 0, I$ omit the terms$\underline{\quad\;}$;

for $j = 0, J$ omit the terms$\utilde{\quad\;}$.

[Note: The points with $i = 0, I$ and $j = 0, J$ (together!) may be bypassed.]

Alternatively, in place of (D.3):

$$\epsilon = \operatorname{Sgn} \alpha, \ \eta = \operatorname{Sgn} \beta \quad .$$

$$\begin{aligned} \overset{h+1}{\rho}_{ij} = {} & \overset{h}{\rho}_{ij} + \underline{(\overset{h}{\rho}_{i+\epsilon j} - \overset{h}{\rho}_{ij}) \mid \alpha \mid} + \utilde{(\overset{h}{\rho}_{ij+\eta}\, \overset{h}{\rho}_{ij}) \mid \beta \mid}\\ & + \utilde{\underline{(\overset{h}{\rho}_{i+\epsilon j+\eta} - \overset{h}{\rho}_{i+\epsilon j} - \overset{h}{\rho}_{ij+\eta} + \overset{h}{\rho}_{ij}) \mid \alpha \mid \mid \beta \mid}} \end{aligned} \tag{D'.3}$$

for $i \neq 0, I$ and $J \neq 0, J$;

for $i = 0, J$ omit the terms$\underline{\quad\;}$;

for $j = 0, J$ omit the terms$\utilde{\quad\;}$.

[Note: The points with $i = 0, I$ and $j = 0, I$ (together!) may be bypassed.]

APPENDIX II

1. The purpose of this paper is to find a rapidly converging iterative method for the solution of linear equation systems, and quite particularly of those which arise from the difference equation treatment of partial differential equations of the elliptic type [2nd order, s ($= 2, 3, \cdots$) variables]. Sections **2–6** are introductory. The method will be described and discussed in Sections **7–13**. The application to the (elliptic) differential equation case will be made in Sections **14–15**. Some comparisons will be made. The results are summarized in Sections **11**, **13**, **15**.

2. Consider a system of n linear equations in n variables, written vectorially:

$$A\boldsymbol{\xi} = \boldsymbol{\alpha}. \tag{1}$$

Here $\boldsymbol{\alpha}$ is a known nth order vector, A a known nth order matrix, $\boldsymbol{\xi}$ the unknown nth order vector. In order that the problem be meaningful, A must be non-singular. This will be assumed.

An iterative method is based on a correction step, which replaces a $\boldsymbol{\xi}$, that may not solve (1), by a $\boldsymbol{\xi}^1$, that, in some suitable sense, should more nearly solve (1). This correction step should be a linear operation F applied to the two nth order vectors $\boldsymbol{\xi}$, $\boldsymbol{\alpha}$, i.e., to the $2n$th order vector $\{\boldsymbol{\xi}, \boldsymbol{\alpha}\}$. It then produces the nth order vector $\boldsymbol{\xi}^1$:

$$\boldsymbol{\xi}^1 = F\{\boldsymbol{\xi}, \boldsymbol{\alpha}\}. \tag{2}$$

It is convenient, to put in place of the nth order vector $\boldsymbol{\xi}^1$ again a $2n$th order vector, namely $\{\boldsymbol{\xi}^1, \boldsymbol{\alpha}\}$. In this case let us write E in place of F:

$$\{\boldsymbol{\xi}^1, \boldsymbol{\alpha}\} = E\{\boldsymbol{\xi}, \boldsymbol{\alpha}\}. \tag{3}$$

Thus E is a $2n$th order matrix.

Note that the linearity of F means that it can be written as follows:

$$F\{\boldsymbol{\xi}, \boldsymbol{\alpha}\} = G\boldsymbol{\xi} + H\boldsymbol{\alpha}, \tag{4}$$

where G, H are nth order matrices.

(2), (3), (4) mean that the $2n$th order matrix E can be written as a 2nd order hypermatrix of nth order matrices, as follows:

$$E = \left(\begin{array}{c|c} G & H \\ \hline O & I \end{array}\right). \tag{5}$$

Here, O, I are the (nth order) zero and unit matrix, as usual.

3. A minimum requirement to be imposed on a correction step in the sense of **2** is this: If $\boldsymbol{\xi}^*$ is the solution of (1), then the correction should leave $\boldsymbol{\xi} = \boldsymbol{\xi}^*$ unchanged, i.e., produce $\boldsymbol{\xi}^1 = \boldsymbol{\xi}(= \boldsymbol{\xi}^*)$. This is the "weak" condition. A reasonable further requirement is that if $\boldsymbol{\xi}$ is not a solution of (1) ($\boldsymbol{\xi}^1 \neq \boldsymbol{\xi}^*$), then the correction should change $\boldsymbol{\xi}$, i.e., produce a $\boldsymbol{\xi}^1 \neq \boldsymbol{\xi}$. This is the "strong" condition. That is, the weak (strong) condition requires that $\boldsymbol{\xi} = \boldsymbol{\xi}^*$ be sufficient (necessary and sufficient) for $\boldsymbol{\xi}^1 = \boldsymbol{\xi}$.

By (2), (4) $\boldsymbol{\xi}^1 = \boldsymbol{\xi}$ means

$$(I - G)\boldsymbol{\xi} = H\boldsymbol{\alpha}. \tag{6}$$

The weak condition requires, that (1) imply (6), i.e., that always

$$(I - G)\boldsymbol{\xi} = HA\boldsymbol{\xi},$$

i.e.,

$$\begin{aligned} I - G &= HA, \\ G &= I - HA. \end{aligned} \tag{7}$$

The strong condition requires, in addition to this, that (6) imply (1), i.e., in view of (7), that

$$HA\boldsymbol{\xi} = H\boldsymbol{\alpha}$$

imply

$$A\boldsymbol{\xi} = \boldsymbol{\alpha}.$$

This means obviously that

$$H \text{ is non-singular.} \tag{8a}$$

Equivalently:

$$0 \text{ is not a characteristic root of } H. \tag{8b}$$

Since A is non-singular, non-singularity of H is equivalent to that of HA, i.e., [by (7)] of $I - G$; i.e., equivalent to this: 0 is not a characteristic root of $I - G$, or equivalently:

$$1 \text{ is not a characteristic root of } G. \tag{8c}$$

To begin with, we will only stipulate the weak condition, i.e., (7).

4. The ordinary iterative procedure consists of repeating the basic step (2) successively, and to expect that the sequence so generated will converge to the solution $\boldsymbol{\xi}^*$ of (1), irrespective of the starting point $\boldsymbol{\xi}$.

Disregarding for the moment the question of convergence, the sequence in question, $\boldsymbol{\xi}^0, \boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \cdots$, is defined by

$$\left.\begin{aligned} \boldsymbol{\xi}^0 &= \boldsymbol{\xi}, \\ \boldsymbol{\xi}^{k+1} &= F\{\boldsymbol{\xi}^k, \boldsymbol{\alpha}\} \qquad (k = 0, 1, 2, \cdots). \end{aligned}\right\} \tag{9}$$

In view of (2), (3), the second equation of (9) can be written

$$\{\boldsymbol{\xi}^{k+1}, \boldsymbol{\alpha}\} = E\{\boldsymbol{\xi}^k, \boldsymbol{\alpha}\} \qquad (k = 0, 1, 2, \cdots),$$

and hence (9) is equivalent to

$$\{\boldsymbol{\xi}^{k+1}, \boldsymbol{\alpha}\} = E^{k+1}\{\boldsymbol{\xi}, \boldsymbol{\alpha}\} \qquad (k = 0, 1, 2, \cdots). \tag{10}$$

We know that for $\boldsymbol{\xi} = \boldsymbol{\xi}^*$ [$\boldsymbol{\xi}^*$ the solution of (1), cf. 3] all $\boldsymbol{\xi}^k = \boldsymbol{\xi}^*$, hence (10) gives

$$\{\boldsymbol{\xi}^*, \boldsymbol{\alpha}\} = E^k\{\boldsymbol{\xi}^*, \boldsymbol{\alpha}\}.$$

Hence (10) is equivalent to what is obtained by subtracting this equation from it, i.e., to

$$\{\boldsymbol{\xi}^k - \boldsymbol{\xi}^*, 0\} = E^k\{\boldsymbol{\xi} - \boldsymbol{\xi}^*, 0\},$$

i.e., in view of (5) to

$$\boldsymbol{\xi}^k - \boldsymbol{\xi}^* = G^k(\boldsymbol{\xi} - \boldsymbol{\xi}^*) \qquad (k = 0, 1, 2, \cdots). \tag{11}$$

Note that (10) is an effective calculational procedure, while (11) is not, since it contains the unknown $\boldsymbol{\xi}^*$; however, some proofs and evaluations can be more advantageously based on (11).

It is well known that frequently the convergence properties of a sequence can be significantly improved by replacing each element of the sequence by a suitable mean of itself and the preceding elements of the sequence. In this sense, one might replace the sequence $\boldsymbol{\xi}^0, \boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \cdots$ by a sequence $\mathfrak{n}^0, \mathfrak{n}^1, \mathfrak{n}^2, \cdots$, where

$$\mathfrak{n}^k = \sum_{l=0}^{k} a_{kl}\boldsymbol{\xi}^l \qquad (k = 0, 1, 2, \cdots), \tag{12}$$

with a suitable set of coefficients a_{kl}. The characterization of the $\boldsymbol{\eta}^k$ as means (of the $\boldsymbol{\xi}^k$) makes it natural to require

$$\sum_{l=0}^{k} a_{kl} = 1 \qquad (k = 0, 1, 2, \cdots). \tag{13}$$

This condition can also be obtained from the natural requirement that for $\boldsymbol{\xi} = \boldsymbol{\xi}^*$, when all $\boldsymbol{\xi}^k = \boldsymbol{\xi}^*$, there shall also be all $\boldsymbol{\eta}^k = \boldsymbol{\xi}^*$. At any rate, we stipulate (13). The characterization of the $\boldsymbol{\eta}^k$ as means might also suggest the requirement that all $a_{kl} \geqq 0$, but we will not impose it; indeed, the choice that we will later make, and that seems to be particularly favorable, will violate this condition [cf. Sections **7**–**9**, in particular (53)].

Instead of working with the coefficients a_{kl} themselves, we can also work with the corresponding polynomials

$$P_k(Z) = \sum_{l=0}^{k} a_{kl}\, Z^l \qquad (k = 0, 1, 2, \cdots). \tag{14}$$

Then (13) becomes

$$P_k(1) = 1. \tag{15}$$

Thus $P_k(Z)$ is a kth order polynomial fulfilling (15), and (so far) subject to no other restrictions.

Now (12) becomes, using (10) [and (15)],

$$\{\boldsymbol{\eta}^k, \boldsymbol{\alpha}\} = P_k(E)\,\{\boldsymbol{\xi}, \boldsymbol{\alpha}\}, \tag{16}$$

or equivalently, using (11),

$$\boldsymbol{\eta}^k - \boldsymbol{\xi}^* = P_k(G)\,(\boldsymbol{\xi} - \boldsymbol{\xi}^*). \tag{17}$$

The relationship between (16), (17) is similar to that between (10), (11), as discussed immediately after (11).

The broader convergence problem for the iterative-and-mean procedure is this: Choose the a_{kl}, i.e., the sequence $[P_0(Z), P_1(Z), P_2(Z), \cdots]$, [$P_k(Z)$ a kth order polynomial fulfilling (15), cf. above], so that

$$\lim_{k\to\infty} \boldsymbol{\eta}^k = \boldsymbol{\xi}^* \tag{18}$$

for all choices of the starting point $\boldsymbol{\xi}$. (18) can be also be written like this:

$$\lim_{k\to\infty} D(\boldsymbol{\eta}^k - \boldsymbol{\xi}^*) = 0, \tag{19}$$

where $D(\boldsymbol{\xi})$ is any norm in the space of all nth order vectors $\boldsymbol{\xi}$ (i.e., in n-dimensional Euclidean space), which is equivalent to the ordinary (Euclidean) topology of that space. We will make a specific choice of $D(\boldsymbol{\xi})$ soon: Section **7** (30).

The ordinary iteration convergence problem (without means, cf. the beginning of **4**) corresponds to the choice $P_k(Z) \equiv Z^k$ $(k = 0, 1, 2, \cdots)$ for the sequence $[P_0(Z), P_1(Z), P_2(Z), \cdots]$.

5. We will now consider the broad convergence problem (iterative-and-mean procedure, cf. **4** above) in more specific detail.

(19) works with

$$d_k = D(\mathbf{n}^k - \boldsymbol{\xi}^*) \qquad (k = 0, 1, 2, \cdots), \tag{20}$$

and its requirement is

$$\lim_{k\to\infty} d_k = 0 \qquad (\text{for all } \boldsymbol{\xi}). \tag{21}$$

Hence the relevant quantity is d_k, and we must concentrate on estimating its size (for all $\boldsymbol{\xi}$).

Combining (20) and (17) gives

$$d_k = D[P_k(G)\omega], \tag{22}$$

where

$$\omega = \boldsymbol{\xi} - \boldsymbol{\xi}^*.$$

(22), (21) show that the convergence problem is actually one of the convergence of the matrices $P_k(G)(k \to \infty)$ to zero.

Thus the problem presents itself in this form: Given a matrix G, what conditions must a sequence of polynomials $[P_0(Z), P_1(Z), P_2(Z), \cdots]$ fulfill so as to have

$$\lim_{k\to\infty} P_k(G) = 0. \tag{23}$$

The answer is well known: Let λ_i ($i = 1, \cdots, \mu$, of course $\mu \leqq n$) be the characteristic roots of G, and let e_i ($= 1, 2, \cdots$) be the order of the elementary divisor of G that corresponds to λ_i. Denote the ρth derivative of $P_k(Z)$ by $P_k^{(\rho)}(Z)$. Then the necessary and sufficient condition for the validity of (23) is this:

$$\lim_{k\to\infty} P_k^{(\rho)}(\lambda_i) = 0 \tag{24}$$

for all those combinations $i(= 1, \cdots, \mu)$, $\rho\ (= 0,1, \cdots)$ for which $e_i > \rho$. When all $e_i = 1$, then (24) becomes simply

$$\lim_{k\to\infty} P_k(\lambda_i) = 0 \qquad (i = 1, \cdots, \mu). \tag{25}$$

For a Hermitian G, in particular, this is always the case.

We saw in **4**, that the $P_k(Z)$ are subject to the condition (15): $P_k(1) = 1$. Hence (24), i.e., (23), is unfulfillable if some $\lambda_i = 1$, i.e., if 1 is a characteristic root of G. In other words: The condition (8c), i.e., the strong condition of **3**, is reimposed for this reason. [This could have been seen directly too: If that condition fails, then for some $\boldsymbol{\xi} \neq \boldsymbol{\xi}^*$ there is $\boldsymbol{\xi}^1 = \boldsymbol{\xi}$, hence all $\boldsymbol{\xi}^k = \boldsymbol{\xi}$, hence all $\mathbf{n}^k = \boldsymbol{\xi}$, and so $\lim_{k\to\infty} \mathbf{n}^k = \boldsymbol{\xi} \neq \boldsymbol{\xi}^*$, contradicting (18).]

If, on the other hand, (8c), i.e., the strong condition in **3**, holds, i.e., if all $\lambda_i \neq 1$, then it is not difficult to see that (24) and (15) are compatible. Indeed, even a fixed $P(Z)$ [for all $P_k(Z)$ with $k \geqq$ the precise order of $P(Z)$, which is $\sum_i e^i$, cf. below] will do:

$$P(Z) \equiv c\prod_i (Z - \lambda_i)^{e_i},$$

with c determined from (15), meets all requirements. This, however, is of small practical importance, since the λ_i may not be known, and the above expression for $P(Z)$ may in any case be too complicated for actual evaluation. If it is only known that the λ_i lie in the interior of a certain (bounded and closed) domain Λ, then a sequence $[P_0(Z), P_1(Z), P_2(Z), \cdots]$ of the desired kind can still be specified, if (and only if) Λ does not separate 1 from ∞. We will, however, not go here into this matter any further.

6. The ordinary iterative procedure corresponds, as we observed at the end of **4**, to the choice $P_k(Z) \equiv Z^k$. Hence $P_k^{(\rho)}(Z) \equiv k(k-1) \cdots (k-\rho+1)Z^{k-\rho}$. Therefore the convergence criterion (24) requires precisely, that all $|\lambda_i| < 1$. We state this explicitly:

(26) The ordinary iterative procedure converges (cf. the beginning of 4) if and only if $|\lambda| < 1$ for all characteristic values λ of G.

As we saw in the last part of **5**, this condition is by no means necessary for the convergence of some suitable iterative-and-mean procedure. We will nevertheless limit ourselves to this case:

$$|\lambda| < 1 \quad \text{for all characteristic roots } \lambda \text{ of } G. \tag{27}$$

In addition, we will assume that G is Hermitian, because this covers certain important applications, and permits the employment of some rather effective methods. We restate this:

$$G \text{ is Hermitian.} \tag{28}$$

(28) implies that all characteristic roots (or characteristic values) of G are real. Hence (27) becomes this:

$$-1 < \lambda < 1 \quad \text{for all characteristic values } \lambda \text{ of } G. \tag{29}$$

Under these conditions the ordinary iterative procedure, i.e., the choice (20), $P_k(Z) \equiv Z^k$ (cf. above), is adequate, i.e., it guarantees convergence. We wish, however, to determine that iterative-and-mean procedure, i.e., that sequence $[P_0(Z), P_1(Z), P_2(Z), \cdots]$ for which this convergence is (uniformly) fastest. We will, therefore reconsider the convergence problem under this aspect, subject to the restrictions (28), (29).

7. We want to choose the sequence $[P_0(Z), P_1(Z), P_2(Z), \cdots]$ so as to obtain the uniformly fastest possible convergence. This convergence is to be taken in the sense of (21), i.e., we want to make for each k the d_k of (20) as small as possible. {By (22) this means that we want to make $D[P_k(G)\omega]$ as small as possible.} This should be true, in some suitable sense, uniformly—i.e., uniformly in the variables of (22). These variables are [since k is given and $P_k(Z)$ is being looked for] G and $\omega(= \xi - \xi^*)$. Let us therefore examine the meaning of uniformity with respect to G and ω.

First, since we are now dealing with a situation in which a Hermitian matrix, G, occupies a central role, it is reasonable to prescribe that the norm $D(\xi)$ be the Euclidean norm

$$D(\xi) = \sqrt{\sum_{i=1}^{n} |\xi_i|^2} \tag{30}$$

($\xi_1, \cdots, \xi_n$ are the [complex, numerical] components of the [nth order] vector ξ). (Cf. the remark following (19) in **4**)

It is also useful to introduce, for a general (nth order) matrix K, the concepts of the "upper bound" $| K |_u$ and of the "lower bound" $| K |_l$:

$$(31a) \qquad | K |_u = \operatorname*{Max}_{\xi \neq 0} \frac{D(K\xi)}{D(\xi)} = \text{Min [C such that } D(K\xi) \leqq CD(\xi)] ,$$

$$(31b) \qquad | K |_l = \operatorname*{Min}_{\xi \neq 0} \frac{D(K\xi)}{D(\xi)} = \text{Max [C such that } D(K\xi) \geqq CD(\xi)] .$$

Second, let us consider the variability of G. G is subject to the requirements (28), (29), i.e., it must be Hermitian and fulfill (29). It is clearly logical to make the requirement (29) with uniformity, i.e., to require for a suitable $\epsilon > 0$ that

$$(32a) \qquad -(1 - \epsilon) \leqq \lambda \leqq 1 - \epsilon$$

for all characteristic values λ of G, or equivalently

$$(32b) \qquad | G |_u \leqq 1 - \epsilon.$$

Third, let us consider the variability of ω. We want to make $D[P_k(G)\omega]$ as small as possible (cf. above). $\omega = 0$ is uninteresting, and it is plausible that we should want to make the ratio $D[P_k(G)\omega]/D(\omega)$ uniformly small for all ω, i.e., to make

$$\operatorname*{Max}_{\omega \neq 0} \frac{D[P_k(G)\omega]}{D(\omega)}$$

small. This means, by (31a), that we want to make $| P_k(G) |_u$ small.

Combining the second and the third remarks, we see that we want to make $| P_k(G) |_u$ uniformly small for all Hermitian G that fulfill (32a) [i.e., (32b)]. That is, we want to minimize

$$(33) \qquad \bar{d}_k = \operatorname*{Max}_{\substack{G \text{ Hermitian} \\ |G|_u \leqq 1-\epsilon}} (| P_k(G) |_u) = \text{Min} \{C \text{ such that for all } \omega \text{ and all Hermitian } G \text{ with } | G |_u \leqq 1 - \epsilon, D[P_k(G)\omega] \leqq CD(\omega)\}.$$

Now $| P_k(G) |_u$ is the maximum $P_k(\lambda)$, where λ runs over all characteristic values of G. In view of the equivalence of (32a) and (32b), the precise limitation on these λ is $-(1 - \epsilon) \leqq \lambda \leqq 1 - \epsilon$. Hence the first part of (33) can be rewritten

$$(34) \qquad \bar{d}_k = \operatorname*{Max}_{-(1-\epsilon) \leqq \lambda \leqq (1-\epsilon)} (| P_k(\lambda) |).$$

Thus we are looking for that kth order polynomial $P_k(Z)$, fulfilling (15), $P_k(1) = 1$, for which

$$\operatorname*{Max}_{-(1-\epsilon) \leqq Z \leqq 1-\epsilon} (| P_k(Z) |)$$

is minimal. Equivalently: We are looking for that kth order polynomial $Q_k(Z)$, fulfilling $| Q_k(Z) | \leqq 1$ for all Z in $-(1 - \epsilon) \leqq Z \leqq (1 - \epsilon)$, for which $Q_k(1)$ is maximal. Indeed, for this $Q_k(1)$ clearly

$$\operatorname*{Max}_{-(1-\epsilon) \leqq Z \leqq 1-\epsilon} (| Q_k(Z) |) = 1,$$

and

$$P_k(Z) \equiv \frac{Q_k(Z)}{Q_k(1)}, \tag{35}$$

$$\operatorname*{Max}_{-(1-\epsilon)\leqq Z\leqq (1-\epsilon)} (\,|\,P_k(Z)\,|\,) = \frac{1}{Q_k(1)}. \tag{36}$$

Again equivalently: We are looking for that kth order polynomial $R_k(Z)$, fulfilling $|\,R(Z)\,| \leqq 1$ for all Z in $-1 \leqq Z \leqq 1$, for which $R_k[1/(1-\epsilon)]$ is maximal. Clearly

$$Q_k(Z) \equiv R_k\left(\frac{Z}{1-\epsilon}\right). \tag{37}$$

8. The last problem in **7** [the one relative to $R_k(Z)$] is classical. It has been solved by Chebyshev [2]. The $R_k(Z)$ in question is the kth Chebyshev-polynomial, defined by

$$R_k\,(\cos u) \equiv \cos\,(ku). \tag{38}$$

It is clear from (38) that $R_k(Z)$ is the kth order polynomial, and that $-1 \leqq Z \leqq 1$ implies $|\,R_k(Z)\,| \leqq 1$, as desired. Putting $u = iv$ gives $R_k(Ch\ v) = Ch(kv)$, putting $e^v = x$ gives $R_k[\frac{1}{2}(x + x^{-1})] = \frac{1}{2}(x^k + x^{-k})$, and putting $x = Z + \sqrt{Z^2 - 1}$ gives

$$R_k(Z) \equiv \tfrac{1}{2}[(Z + \sqrt{Z^2-1})^k + (Z + \sqrt{Z^2-1})^{-k}]. \tag{39}$$

Now putting $Z = 1/(1-\epsilon)$ gives

$$R_k\left(\frac{1}{1-\epsilon}\right) = \frac{1}{2}\left\{\left[\frac{1+\sqrt{(2-\epsilon)\epsilon}}{1-\epsilon}\right]^k + \left[\frac{1+\sqrt{(2-\epsilon)\epsilon}}{1-\epsilon}\right]^{-k}\right\}. \tag{40}$$

Combining (37) with (35), (36) gives:

$$P_k(Z) \equiv \frac{R_k\left(\dfrac{Z}{1-\epsilon}\right)}{R_k\left(\dfrac{1}{1-\epsilon}\right)}, \tag{41}$$

$$\operatorname*{Max}_{-(1-\epsilon)\leqq Z\leqq 1-\epsilon} (|P_k(Z)|) = \frac{1}{R_k\left(\dfrac{1}{1-\epsilon}\right)}. \tag{42}$$

It is worthwhile to compare the efficiency of this scheme with that of the ordinary iterative procedure (without means), i.e., with the choice $P_k(Z) \equiv Z^k$ (cf. the end of **4** and the beginning of **6**).

Consider first the present choice for $P_k(Z)$ [i.e., (41)]. The logarithm of the first term in the bracket on the right hand side of (40) is

$$\ln\left(\frac{1+\sqrt{(2-\epsilon)\epsilon}}{1-\epsilon}\right)\cdot k,$$

i.e., for $\epsilon \ll 1$ it is $\sim \sqrt{2\epsilon}\cdot k$. The logarithm of the second term is correspondingly $\sim -\sqrt{2\epsilon}\cdot k$. Assume furthermore $\sqrt{2\epsilon}\cdot k \gg 1$, then the first term is dominant, i.e.,

$$R_k\left(\frac{1}{1-\epsilon}\right) \sim \tfrac{1}{2}e^{h_1}$$

with $h_1 \sim \sqrt{2\epsilon}\cdot k$, and so by (42)

$$\text{(43a)} \qquad \underset{-(1-\epsilon) \leqq Z \leqq 1-\epsilon}{\text{Max}} \left(| P_k(Z) | \right) = 2e^{-h_1}$$

with $h_1 \sim \sqrt{2\epsilon}\cdot k$, if $\epsilon \ll 1$, $\sqrt{2\epsilon}\cdot k \gg 1$.

Consider next the choice $P_k(Z) \equiv Z^k$. Then clearly

$$\underset{-(1-\epsilon) \leqq Z \leqq 1-\epsilon}{\text{Max}} \left(| P_k(Z) | \right) = (1-\epsilon)^k.$$

The logarithm of the right hand side is $\ln(1-\epsilon)\cdot k$, i.e., for $\epsilon \ll 1$ it is $\sim \epsilon\cdot k$. Hence in this case

$$\text{(43b)} \qquad \underset{-(1-\epsilon) \leqq Z \leqq 1-\epsilon}{\text{Max}} \left(| P_k(Z) | \right) = e^{-h_2}$$

with $h_2 \sim \epsilon\cdot k$, if $\epsilon \ll 1$.

Comparing (43a) and (43b), and remembering (34) shows that the speed of uniform convergence, i.e., the speed of decrease of $\bar{d}_k$, compares as follows for the choices of $P_k(Z)$ under consideration—namely, the "optimum" choice of $P_k(Z)$ [i.e., (41)], and the "ordinary" (no means!) choice of $P_k(Z)$ (i.e., $\equiv Z^k$): In the first case the increase of k that e^{-1}-folds $\bar{d}_k$ (asymptotically!) is $\Delta k = 1/\sqrt{2\epsilon}$, in the second case that increase is $\Delta k = 1/\epsilon$. Thus the first choice accelerates the convergence over the second choice in the ratio $\sqrt{2\epsilon}:\epsilon = \sqrt{2/\epsilon}$.

9. Let us now return to the definition (38) of $R_k(Z)$ [on which the "optimum" definition (41) of $P_k(Z)$ is based]. (38) is transcendental, the equivalent (39) is irrational. It is desirable to replace these by a rational definition. Such a definition obtains, in the form of a two-step recursion, from the identity

$$\cos\,[(k+1)u] + \cos\,[(k-1)u] \equiv 2\cos u \cos\,(ku).$$

In view of (38) this gives

$$R_{k+1}(Z) + R_{k-1}(Z) \equiv 2ZR_k(Z),$$

i.e.,

$$\text{(44)} \qquad R_{k+1}(Z) \equiv 2ZR_k(Z) - R_{k-1}(Z) \qquad (k = 1, 2, \cdots).$$

This relation, together with the "starting conditions"

$$\text{(45)} \qquad R_0(Z) \equiv 1, \qquad R_1(Z) \equiv Z,$$

defines the $R_k(Z)$ completely.

Now (41) permits us to pass to $P_k(Z)$. Then (44) becomes

$$P_{k+1}(Z) \equiv 2\,\frac{Z}{1-\epsilon}\,\frac{R_k\left(\dfrac{1}{1-\epsilon}\right)}{R_{k+1}\left(\dfrac{1}{1-\epsilon}\right)}\,P_k(Z) - \frac{R_{k-1}\left(\dfrac{1}{1-\epsilon}\right)}{R_{k+1}\left(\dfrac{1}{1-\epsilon}\right)}\,P_{k-1}(Z),$$

i.e.,

$$(46)\qquad P_{k+1}(Z) \equiv 2Z\,\frac{a_{k+1}}{1-\epsilon}\,P_k(Z) - a_{k+1}\,a_k\,P_{k-1}(Z),$$

where

$$(47)\qquad a_l = \frac{R_{l-1}\left(\dfrac{1}{1-\epsilon}\right)}{R_l\left(\dfrac{1}{1-\epsilon}\right)}.$$

Putting $Z = 1/(1-\epsilon)$ in (44) and dividing by $R_k[1/(1-\epsilon)]$ gives

$$(48)\qquad \frac{1}{a_{k+1}} = \frac{2}{1-\epsilon} - a_k\,.$$

Through (41), (45) becomes

$$(49)\qquad P_0(Z) \equiv 1, \qquad P_1(Z) \equiv Z.$$

Also, (45) gives

$$(50)\qquad a_1 = 1 - \epsilon.$$

It is convenient to introduce

$$(51)\qquad b_l = \frac{a_l}{1-\epsilon}\,.$$

Then (50), (48) give

$$(52a)\qquad b_1 = 1,$$

$$(52b)\qquad b_{k+1} = \frac{1}{2-(1-\epsilon)^2 b_k} \qquad (k = 1, 2, \cdots).$$

Next, (46) gives

$$P_{k+1}(Z) \equiv 2b_{k+1}Z\,P_k(Z) - (1-\epsilon)^2\,b_{k+1}\,b_k\,P_{k-1}(Z).$$

Owing to (52b)

$$(1-\epsilon)^2\,b_{k+1}\,b_k = 2b_{k+1} - 1,$$

hence the above equation can also be written like this:

$$(53)\qquad P_{k+1}(Z) = 2b_{k+1}\,[ZP_k(Z) - P_{k-1}(Z)] + P_{k-1}(Z) \qquad (k = 1, 2, \cdots).$$

Finally by (34), (42)

$$\bar{d}_k = \frac{1}{R_k\left(\dfrac{1}{1-\epsilon}\right)},$$

hence by (45), (47)

$$\bar{d}_k = a_1 \cdots a_k\,,$$

and so by (51)

$$\bar{d}_k = (1 - \epsilon)^k \, b_1 \cdots b_k \,. \tag{54}$$

10. We can now pass from the $P_k(Z)$ to the $\mathbf{n}^{(k)}$, of course with the help of (16). We replace Z by the $2n$th order matrix E in both equations of (49) as well as in (53). Thus in all three equations both sides become $2n$th order matrices. We apply these to the $2n$th order vector $\{\boldsymbol{\xi}, \boldsymbol{\alpha}\}$. In this way three equations obtain, each one having $2n$th order vectors on both sides. These are as follows:

From the first equation of (49), using (16):

$$\{\mathbf{n}^0, \boldsymbol{\alpha}\} = \{\boldsymbol{\xi}, \boldsymbol{\alpha}\},$$

i.e.,

$$\mathbf{n}^0 = \boldsymbol{\xi}. \tag{55}$$

From the second equation of (49), using (16):

$$\{\mathbf{n}^1, \boldsymbol{\alpha}\} = E\{\boldsymbol{\xi}, \boldsymbol{\alpha}\},$$

i.e., recalling (2), (3):

$$\mathbf{n}^1 = F\{\boldsymbol{\xi}, \boldsymbol{\alpha}\}. \tag{56}$$

From (53), using (16):

$$\{\mathbf{n}^{k+1}, \boldsymbol{\alpha}\} = 2b_{k+1}(E\{\mathbf{n}^k, \boldsymbol{\alpha}\} - \{\mathbf{n}^{k-1}, \boldsymbol{\alpha}\}) + \{\mathbf{n}^{k-1}, \boldsymbol{\alpha}\},$$

i.e., again recalling (2), (3):

$$\begin{aligned}\{\mathbf{n}^{k+1}, \boldsymbol{\alpha}\} &= 2b_{k+1}[(F\{\mathbf{n}^k, \boldsymbol{\alpha}\}, \boldsymbol{\alpha}) - \{\mathbf{n}^{k-1}, \boldsymbol{\alpha}\}] + \{\mathbf{n}^{k-1}, \boldsymbol{\alpha}\} \\ &= 2b_{k+1}(F\{\mathbf{n}^k, \boldsymbol{\alpha}\} - \mathbf{n}^{k-1}) + \{\mathbf{n}^{k-1}, \boldsymbol{\alpha}\},\end{aligned}$$

i.e.,

$$\mathbf{n}^{k+1} = 2b_{k+1}(F\{\mathbf{n}^k, \boldsymbol{\alpha}\} - \mathbf{n}^{k-1}) + \mathbf{n}^{k-1}, \qquad (k = 1, 2, \cdots). \tag{57}$$

11. We have obtained an inductive definition of the sequence $\mathbf{n}^0, \mathbf{n}^1, \mathbf{n}^2, \cdots$. This is based on another, inductively defined, (numerical) sequence $b_1, b_2, \cdots$. Actually the two inductions can proceed concurrently. We will now restate these.

The b_k induction is given by (52a), (52b):

$$b_1 = 1, \tag{Ia}$$

$$b_{k+1} = \frac{1}{2 - (1 - \epsilon)^2 b_k} \qquad (k = 1, 2, \cdots). \tag{Ib}$$

The $\mathbf{n}^k$ induction is given by (55), (56), (57):

$$\mathbf{n}^0 = \boldsymbol{\xi}, \tag{IIa}$$

$$\mathbf{n}^1 = F\{\boldsymbol{\xi}, \boldsymbol{\alpha}\}, \tag{IIb}$$

$$\mathbf{n}^{k+1} = 2b_{k+1}(F\{\mathbf{n}^k, \boldsymbol{\alpha}\} - \mathbf{n}^{k-1}) + \mathbf{n}^{k-1} \qquad (k = 1, 2, \cdots). \tag{IIc}$$

We also restate the formula (54) for $\bar{d}_k$:

$$\bar{d}_k = (1 - \epsilon)^k b_1 \cdots b_k \,. \tag{III}$$

This can be expressed inductively:

$$\text{(IIIa)} \qquad \bar{d}_0 = 1,$$

$$\text{(IIIa)} \qquad \bar{d}_{k+1} = (1-\epsilon)b_{k+1}\bar{d}_k \qquad (k = 0, 1, 2, \cdots).$$

It is worthwhile to compare this process, and in particular its central piece (II) (which produces the sequence $\mathbf{n}^0, \mathbf{n}^1, \mathbf{n}^2, \cdots$), with the ordinary iterative process, i.e., with (9) (which produces the sequence $\boldsymbol{\xi}^0, \boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \cdots$). (II) and (9) give the same $\mathbf{n}^k$ and $\boldsymbol{\xi}^k$ for $k = 0, 1$, but they differ for $k = 2, 3, \cdots$, i.e., for $\mathbf{n}^{k+1}$ and $\boldsymbol{\xi}^{k+1}$ for $k = 1, 2, \cdots$. Even here the first step in forming $\mathbf{n}^{k+1}$ is the same as the (only) step in forming $\boldsymbol{\xi}^{k+1}$, the application of the original correction step $F\{\cdots, \boldsymbol{\alpha}\}$ (cf. 2). In forming $\mathbf{n}^{k+1}$, however, this is followed by the further step $2b_{k+1}(\cdots - \mathbf{n}^{k-1}) + \mathbf{n}^{k-1}$. This is clearly an extrapolation from $\mathbf{n}^{k-1}$ with the (excess) factor $2b_{k+1} - 1$. Note, that (I) implies $\frac{1}{2} < b_l < 1$ (for all $l = 1, 2, \cdots$), hence $0 < 2b_l - 1 < 1$. Thus the extrapolation (excess) factor lies between 0 and 1.

Now it is by no means unusual that an iterative correction method is improved by combination with extrapolation steps. The noteworthy circumstance is rather, that, in going from $\mathbf{n}^k$ to $\mathbf{n}^{k+1}$, the extrapolation should issue from $\mathbf{n}^{k-1}$. It is also of interest, that a "universal" and "optimum" sequence of extrapolation factors (i.e., the $2b_{k+1} - 1$) could be determined [by (I)].

12. The procedure summarized in **11** is complete, but it is based on the knowledge of a Hermitian G fulfilling (32a) [or equivalently (32b)]. Thus there remains the problem of constructing such a G.

More precisely, we need the F of (2), i.e., the G, H of (4). These are linked by the relation (7), which we restate:

$$\text{(58)} \qquad G = I - HA.$$

A is, of course, given. Thus H is arbitrary, it determines G by (58), and this G must then be Hermitian and fulfilling (32a). These conditions can also be stated in terms of $I - G = HA$: The Hermitian character of G is equivalent to that of HA. (32a) is equivalent to

$$\text{(59)} \qquad \epsilon \leqq \lambda \leqq 2 - \epsilon$$

for all characteristic values of λ of HA.

We repeat: We are looking for an H that makes HA Hermitian and fulfills (59).

We will now describe two procedures that achieve this:

First, put

$$\text{(60)} \qquad H = \alpha A^* \qquad (\alpha > 0).\dagger$$

Then (58) gives

$$\text{(61)} \qquad G = I - \alpha A^* A.$$

$HA = \alpha A^* A$ is obviously Hermitian, it is also positive-definite. Hence the smallest characteristic value of HA is $|\alpha A^* A|_l = \alpha |A^* A|_l = \alpha(|A|_l)^2$, and the largest characteristic value of HA is $|\alpha A^* A|_u = \alpha |A^* A|_u = \alpha(|A|_u)^2$. Consequently

† A^* is the "adjoint" of A, i.e., its complex-conjugate transposed.

(59) means that

$$\alpha(|A|_l)^2 \geqq \epsilon, \tag{62a}$$

$$\alpha(|A|_u)^2 \leqq 2 - \epsilon. \tag{62b}$$

Assume that we know that

$$0 < a \leqq |A|_l \leqq |A|_u \leqq b, \tag{63}$$

i.e., so that a, b are known. Then (62a), (62b) can be guaranteed by prescribing

$$\alpha a^2 = \epsilon, \qquad \alpha b^2 = 2 - \epsilon, \tag{64}$$

i.e.,

$$\alpha = \frac{2}{a^2 + b^2}, \tag{65}$$

$$\epsilon = \frac{2a^2}{a^2 + b^2}. \tag{66}$$

Now put

$$f = \frac{b}{a}. \tag{67}$$

Then (66) becomes

$$\epsilon = \frac{2}{f^2 + 1}. \tag{68}$$

Second, assume that A is Hermitian and positive-definite. In this case put

$$H = \alpha I \qquad (\alpha > 0). \tag{69}$$

Then (58) gives

$$G = I - \alpha A. \tag{70}$$

$HA = \alpha A$ is clearly Hermitian and positive-definite. Hence the smallest characteristic value of HA is $|\alpha A|_l = \alpha |A|_l$, and the largest characteristic value of HA is $|\alpha A|_u = \alpha |A|_u$. Consequently (59) means that

$$\alpha |A|_l \geqq \epsilon, \tag{71a}$$

$$\alpha |A|_u \leqq 2 - \epsilon. \tag{71b}$$

Assuming again the validity of (63), we can guarantee (71a), (71b) by prescribing

$$\alpha a = \epsilon, \qquad \alpha b = 2 - \epsilon, \tag{72}$$

i.e.,

$$\alpha = \frac{2}{a + b}, \tag{73}$$

$$\epsilon = \frac{2a}{a + b}. \tag{74}$$

Using (67) again, (74) becomes

$$\epsilon = \frac{2}{f+1}, \tag{75}$$

13. The results of **12** deserve restatement and some comments. The common assumptions of both parts of **12** are (63), (67):

$$0 < a \leqq | A |_l \leqq | A |_u \leqq b, \tag{IVa}$$

$$f = \frac{b}{a}. \tag{IVb}$$

The result of the first part is contained in (60), (61), (65), (66), (68):

$$H = \alpha A^* \tag{Va}$$

$$G = I - \alpha A^* A, \tag{Vb}$$

$$\alpha = \frac{2}{a^2 + b^2}, \tag{Vc}$$

$$\epsilon = \frac{2a^2}{a^2 + b^2} = \frac{2}{f^2 + 1}, \tag{Vd}$$

A being otherwise unrestricted.

The result of the second part is contained in (69), (70), (73), (74), (75):

$$H = \alpha I, \tag{VIa}$$

$$G = I - \alpha A, \tag{VIb}$$

$$\alpha = \frac{2}{a + b}, \tag{VIc}$$

$$\epsilon = \frac{2a}{a + b} = \frac{2}{f + 1}, \tag{VId}$$

A being assumed Hermitian and positive-definite.

(Va), (Vb) show that the first case is related to the iterative "steepest descent" methods; (VIa), (VIb) show that the second case is related to the iterative "relaxation" methods.

Our derivation makes it plausible why the former are of universal applicability, while the latter are limited to Hermitian and positive-definite matrices—i.e., if the problem arises from the difference equation treatment of partial differential equations of the elliptic type [2nd order, s ($= 2, 3, \cdots$) variables, cf. **1** and again **14**], to the self-adjoint, elliptic case.

In general $f \gg 1$. Then in the first case $\epsilon \sim 2f^{-2}$ [by (Vd)], and in the second case $\epsilon \sim 2f^{-1}$ [by (VId)]. Thus the first case gives a much smaller ϵ than the second case, i.e., in view of the remarks at the end of **8**, a much slower convergence of the iterative process.

This observation illustrates the general experience that whenever relaxation-type procedures are applicable, the convergence is significantly faster than otherwise.

14. We now pass to the consideration of the difference equation system for an elliptic partial differential equation [2nd order, s ($= 2, 3, \cdots$) variables].

Let the partial differential equation be

$$-\sum_{i=1}^{s}\frac{\partial}{\partial x_i}\left(a^i\frac{\partial \xi}{\partial x_i}\right) = \alpha, \tag{76}$$

where $x_1, \cdots, x_s$ are the independent variables, $\xi \equiv \xi(x_1, \cdots, x_s)$ is the dependent variable, and $a^i \equiv a^i(x_1, \cdots, x_s)$, $(i = 1, \cdots, s)$ and $\alpha \equiv \alpha(x_1, \cdots, x_s)$ are known functions of $x_1, \cdots, x_s$. Also

$$0 < \bar{a}^i \leqq a^i(x_1, \cdots, x_s) \leqq \bar{b}^i \qquad (i = 1, \cdots, s), \tag{77}$$

the $\bar{a}^i$, $\bar{b}^i$ $(i = 1, \cdots, s)$ being known constants. Finally the domain of the $x_1, \cdots, x_s$ is

$$0 \leqq x_i \leqq L_i \qquad (i = 1, \cdots, s), \tag{78}$$

and the boundary condition is

$$\xi = 0 \quad \text{for} \quad x_i = 0 \quad \text{or} \quad L_i \qquad (i = 1, \cdots, s). \tag{79}$$

In order to pass to difference equations, we introduce a lattice

$$x_i = \eta_i\, \Delta x_i \left(\Delta x_i = \frac{L_i}{N_i}\right), \tag{80}$$

where in some cases

$$\eta_i = 0, 1 \cdots, N_i - 1, N_i, \tag{80a}$$

in others

$$\eta_i = 1, \cdots, N_i - 1, \tag{80b}$$

and in others again

$$\eta_i = \tfrac{1}{2}, \tfrac{3}{2}, \cdots, N_i - \tfrac{1}{2} \qquad (i = 1, \cdots, s). \tag{80c}$$

Of course, $N_i = 2, 3, \cdots$, and it expresses the fineness of this lattice in the x_i-direction.

We write

$$\xi(x_1, \cdots x_s) = \xi_{\eta_1 \cdots \eta_s}, \tag{81}$$

using (80a). These $\xi_{\eta_1 \cdots \eta_s}$ are the unknowns, but since (79) gives

$$\xi_{\eta_1 \cdots \eta_s} = 0 \quad \text{for} \quad \eta_i = 0, N_i \qquad (i = 1, \cdots, s), \tag{82}$$

the unknown character is actually restricted to (80b).

It is convenient to use with a^i (80b) for the η_j with $j \neq i$, and (80c) for η_i :

$$a^i(x_1, \cdots, x_s) = a^i_{\eta_1 \cdots \eta_s}, \tag{83}$$

and for α (80b) throughout:

$$\alpha(x_1, \cdots, x_s) = \alpha_{\eta_1 \cdots \eta_s}, \tag{84}$$

these being known quantities.

Now (76) is best stated for (80b). It becomes

$$(85)\qquad -\sum_{i=1}^{s}\left(\frac{N_i}{L_i}\right)^2 [a^i_{\eta_1\cdots\eta_i+\frac{1}{2}\cdots\eta_s}(\xi_{\eta_1\cdots\eta_i+1\cdots\eta_s}-\xi_{\eta_1\cdots\eta_i\cdots\eta_s}) \\ - a^i_{\eta_1\cdots\eta_i-\frac{1}{2}\cdots\eta_s}(\xi_{\eta_i\cdots\eta_i\cdots\eta_s}-\xi_{\eta_1\cdots\eta_i-1\cdots\eta_s})] = \alpha_{\eta_1\cdots\eta_s}.$$

We can view (85) as the equivalent of (1), with the following provisos: The complex $\eta_1, \cdots, \eta_s$, according to (80b), stands for the vector-index in (1). Hence the order of the matrix A is

$$(86)\qquad \eta = \prod_i^s (N_i - 1).$$

The $\xi_{\eta_1\cdots\eta_s}$ are therefore the components of the (unknown) vector $\boldsymbol{\xi}$, the $\alpha_{\eta_1\cdots\eta_s}$ are the components of the (known) vector $\boldsymbol{\alpha}$. The left-hand side of (85) then defines the (known) matrix A. Hence

$$(87)\qquad A = \sum_{i=1}^{s}\left(\frac{N_i}{L_i}\right)^2 A_i,$$

where the matrix A_i is defined by

$$(88)\qquad \begin{aligned} A_i\boldsymbol{\xi} &= \boldsymbol{\xi}^+, \\ \xi^+_{\eta_1\cdots\eta_s} &= -a^i_{\eta_1\cdots\eta_i+\frac{1}{2}\cdots\eta_s}(\xi_{\eta_1\cdots\eta_s+1\cdots\eta_s}-\xi_{\eta_1\cdots\eta_i\cdots\eta_s}) \\ &\quad + a^i_{\eta_1\cdots\eta_i-\frac{1}{2}\cdots\eta_s}(\xi_{\eta_1\cdots\eta_i\cdots\eta_s}-\xi_{\eta_1\cdots\eta_i-1\cdots\eta_s}). \end{aligned}$$

Furthermore, clearly

$$(89)\qquad A_i = B_i^*B_i$$

[cf. footnote on page 177, where

$$(90)\qquad \begin{aligned} B_i\boldsymbol{\xi} &= \boldsymbol{\xi}^+, \\ \xi^+_{\eta_1\cdots\eta_s} &= \sqrt{a^i_{\eta_1\cdots\eta_i+\frac{1}{2}\cdots\eta_s}}\,(\xi_{\eta_1\cdots\eta_i+1\cdots\eta_s}-\xi_{\eta_1\cdots\eta_i\cdots\eta_s}). \end{aligned}$$

In order to apply the results of **13**, we now need estimates of $|A|_l$, $|A|_u$, in accordance with (IV) in **13.** From (87)

$$(91)\qquad \left.\begin{aligned} |A|_l &\geqq \sum_{i=1}^{s}\left(\frac{N_i}{L_i}\right)^2 |A_i|_l, \\ |A|_u &\leqq \sum_{i=1}^{s}\left(\frac{N_i}{L_i}\right)^2 |A_i|_u. \end{aligned}\right\}$$

From (89)

$$(92)\qquad \left.\begin{aligned} |A_i|_l &= (|B_i|_l)^2, \\ |A_i|_u &= (|B_i|_u)^2. \end{aligned}\right\}$$

Finally, designate A_i, B_i with $a^i_{\eta_1\cdots\eta_i\cdots\eta_s} = 1$ [cf. the remark preceding (83)!] by A_i^0, B_i^0. Then clearly

$$(93)\qquad \left.\begin{aligned} |B_i|_l &\geqq \sqrt{\bar{a}^i}\,|B_i^0|_l, \\ |B_i|_u &\leqq \sqrt{\bar{b}^i}\,|B_i^0|_u. \end{aligned}\right\}$$

Combining both sides of (93) with (92) gives

$$\left.\begin{aligned} | A_i |_l &\geqq \bar{a}^i \, | A_i^0 |_l , \\ | A_i |_u &\leqq \bar{b}^i \, | A_i^0 |_u , \end{aligned}\right\} \tag{94}$$

and combining (94) with (91) gives

$$\left.\begin{aligned} | A |_l &\geqq \sum_{i=1}^{s} \bar{a}^i \left(\frac{N_i}{L_i}\right)^2 | A_i^0 |_l , \\ | A |_u &\leqq \sum_{i=1}^{s} \bar{b}^i \left(\frac{N_i}{L_i}\right)^2 | A_i^0 |_u . \end{aligned}\right\} \tag{95}$$

Now consider A_i^0. Applying (88) with $a^i_{\eta_1 \cdots \eta_i \cdots \eta_s} = 1$ shows, that the role of the η_j, $j \neq i$, is now irrelevant in determining $| A_i^0 |_l$, $| A_i^0 |_u$. Hence we may write in place of (88)

$$\left.\begin{aligned} A_i^0 \xi &= \xi^+, \\ \xi^+_{\eta_i} &= -\xi_{\eta_i+1} + 2\xi_{\eta_i} - \xi_{\eta_i-1} . \end{aligned}\right\} \tag{96}$$

This operator is Hermitian, its characteristic vectors are the $\xi^{m_i}(m_i = 1, \cdots, N_i - 1)$ with

$$\xi^{m_i}_{\eta_i} = \sin \frac{\pi m_i \, \eta_i}{N_i} , \tag{97}$$

the characteristic value of ξ^{m_i} being

$$\lambda^{\eta_i} = 2 - 2 \cos \frac{\pi m_i}{N_i} = 4 \sin^2 \frac{\pi m_i}{2N_i} . \tag{98}$$

Hence

$$\left.\begin{aligned} | A_i^0 |_l &= \operatorname*{Min}_{m_i} \lambda^{m_i} = \lambda^1 = 4 \sin^2 \frac{\pi}{2N_i} , \\ | A_i^0 |_u &= \operatorname*{Max}_{m_i} \lambda^{m_i} = \lambda^{N_1 - 1} = 4 \cos^2 \frac{\pi}{2N_i} . \end{aligned}\right\} \tag{99}$$

Combining (95) and (99) gives

$$\left.\begin{aligned} | A |_l &\geqq 4 \sum_{i=1}^{s} \bar{a}_i \left(\frac{N_i}{L_i}\right)^2 \sin^2 \frac{\pi}{2N_i} , \\ | A |_u &\leqq 4 \sum_{i=1}^{s} \bar{b}_i \left(\frac{N_i}{L_i}\right)^2 \cos^2 \frac{\pi}{2N_i} . \end{aligned}\right\} \tag{100}$$

Hence we can put in (IVa)

$$\left.\begin{aligned} a &= 4 \sum_{i=1}^{s} \bar{a}_i \left(\frac{N_i}{L_i}\right)^2 \sin^2 \frac{\pi}{2N_i} , \\ b &= 4 \sum_{i=1}^{s} \bar{b}_i \left(\frac{N_i}{L_i}\right)^2 \cos^2 \frac{\pi}{2N_i} . \end{aligned}\right\} \tag{101}$$

Therefore (IVb) gives

$$f = \frac{\sum_{i=1}^{s} \bar{b}_i \left(\frac{N_i}{L_i}\right)^2 \cos^2 \frac{\pi}{2N_i}}{\sum_{i=1}^{s} \bar{a}_i \left(\frac{N_i}{L_i}\right)^2 \sin^2 \frac{\pi}{2N_i}}. \tag{102}$$

If

$$\operatorname*{Max}_{i=1,\cdots,s} \frac{\bar{b}_i}{\bar{a}_i} = \bar{M}, \qquad \operatorname*{Max}_{i=1,\cdots,s} N_i = \bar{N}, \tag{103}$$

Then (102) gives

$$f \leqq \bar{M} \cot^2 \frac{\pi}{2\bar{N}}. \tag{104}$$

and, since

$$\tan \frac{\pi}{2\bar{N}} \geqq \frac{\pi}{2\bar{N}}, \qquad \cot \frac{\pi}{2\bar{N}} \leqq \frac{2\bar{N}}{\pi},$$

a fortiori

$$f \leqq \frac{4}{\pi^2} \bar{M} \bar{N}^2. \tag{105}$$

Now (VId) in **13** gives $\epsilon = 2/(f + 1)$ and the remarks at the end of **8** give for the error-e^{-1}-folding increase of k (asymptotically!)

$$\Delta k \sim \frac{1}{\sqrt{2\epsilon}} = \frac{1}{2}\sqrt{f+1} \sim \frac{1}{2}\sqrt{f},$$

i.e.,

$$\Delta k \sim \frac{1}{2}\sqrt{f}. \tag{106}$$

Hence, in view of (105),

$$\Delta k \lesssim \frac{1}{\pi}\sqrt{\bar{M}}\bar{N}. \tag{107}$$

15. We restate the results of **14.** The elliptic partial differential equation is given in (76), the subsidiary conditions in (77)–(79), the lattice is defined in (80) and (81)–(84), the difference equation system in (85). We do not restate these.

The a, b, f of (IV) in **13** are given in (101), (102):

$$a = 4 \sum_{i=1}^{s} \bar{a}_i \left(\frac{N_i}{L_i}\right)^2 \sin^2 \frac{\pi}{2N_i}, \tag{VIIa}$$

$$b = 4 \sum_{i=1}^{s} \bar{b}_i \left(\frac{N_i}{L_i}\right)^2 \cos^2 \frac{\pi}{2N_i}, \tag{VIIb}$$

$$f = \frac{\sum_{i=1}^{s} \bar{b}_i \left(\frac{N_i}{L_i}\right) \cos^2 \frac{\pi}{2N_i}}{\sum_{i=1}^{s} \bar{a}_i \left(\frac{N_i}{L_i}\right) \sin^2 \frac{\pi}{2N_i}}. \tag{VIIc}$$

If

$$\text{(VIIIa)} \qquad \operatorname*{Max}_{i=1,\cdots,s} \frac{\bar{b}_i}{\bar{a}_i} = \bar{M}, \qquad \operatorname*{Max}_{i=1,\cdots,s} N_i = \bar{N},$$

then

$$\text{(VIIIb)} \qquad f \leqq \bar{M} \cot^2 \frac{\pi}{2\bar{N}} \leqq \frac{4}{\pi^2} \bar{M}\bar{N}^2 .$$

Los Alamos, New Mexico

1. A. Blair, N. Metropolis, J. von Neumann, A. H. Taub & M. Tsingou, *A Study of a Numerical Solution to a Two-dimensional Hydrodynamical Problem*, Los Alamos Report LA-2165, 1958.

2. S. Bernstein, *Leçons sur les Propriétés Extrémales et la Meilleure Approximation des Fonctions Analytiques d'une Variable Réelle*, Gauthier-Villars, Paris, 1926, p. 7–8. Also P. Chebyshev, *Collected Works*, v. 2.

The Determination of the Chebyshev Approximating Polynomial for a Differentiable Function

F. D. Murnaghan and J. W. Wrench, Jr.

If $f(x)$ is continuous over any interval, which we may take, without loss of generality, to be the interval $-1 \leqq x \leqq 1$, there exists a unique polynomial $P_n{}^*(x)$, of given maximum degree n, which is such that the maximum of $|f(x) - P_n{}^*(x)|$ over $-1 \leqq x \leqq 1$ is less than the maximum of $|f(x) - P_n(x)|$ over $-1 \leqq x \leqq 1$, where $P_n(x)$ is any other polynomial of degree not exceeding n. The polynomial $P_n{}^*(x)$ is known as the Chebyshev approximation, of maximum degree n, to $f(x)$ over $-1 \leqq x \leqq 1$. It is characterized by the fact that $f(x) - P_n{}^*(x)$ assumes extreme values at $n + 2$ points, at least, of the interval $-1 \leqq x \leqq 1$, these extreme values being equal in magnitude and alternating in sign [1]. We refer to the points of any such set of $n + 2$ points as critical points, and we denote them by $(x_1{}^*, \cdots, \overset{*}{x}_{n+2})$, where $-1 \leqq x_1{}^* < \cdots < \overset{*}{x}_{n+2} \leqq 1$. Thus, the end points, ± 1, of the interval $-1 \leqq x \leqq 1$ may be critical points, but at least n of the $n + 2$ critical points, namely, $x_2{}^*, \cdots, \overset{*}{x}_{n+1}$, are interior points of this interval.

We assume that $f(x)$ is not only continuous, but also differentiable, over $-1 \leqq x \leqq 1$, and so the derivative of $f(x) - P_n{}^*(x)$ is zero at each of the n points $x_2{}^*, \cdots, \overset{*}{x}_{n+1}$. If the derivative of $f(x) - P_n{}^*(x)$ cannot be zero more than n times, it follows that $x_1{}^* = -1$, $\overset{*}{x}_{n+2} = 1$ and that the derivative of $f(x) - P_n{}^*(x)$ has precisely n zeros that are interior points of the interval $-1 \leqq x \leqq 1$.

The polynomial $P_n{}^*(x)$ is an odd function of x when $f(x)$ is odd, and is an even function of x when $f(x)$ is even. Thus, when $f(x)$ is odd we may take n to be even, the maximum degree of $P_n{}^*(x)$ being $n - 1$, and when $f(x)$ is even we may take n to be odd, the maximum degree of $P_n{}^*(x)$ being again $n - 1$. In these cases the critical points are distributed symmetrically about the mid-point $x = 0$ of the interval $-1 \leqq x \leqq 1$, and we may confine our attention to the part $0 \leqq x \leqq 1$ of this interval. When $f(x)$ is odd, so that n is even, the number of critical points is even and $x = 0$ is not a critical point; on the other hand, when $f(x)$ is even, the number of critical points is odd and $x = 0$ is a critical point. When $f(x)$ is odd, or even, and $x = 1$ is a critical point, we change our notation and denote the positive interior critical points by $x_1{}^* < x_2{}^* < \cdots < x_k{}^*$, where $n = 2k$ in the first case, and $n = 2k + 1$ in the second. For example, when $f(x) = \text{arc tan } x$, $\overset{*}{P}_{2k}(x)$ is an odd polynomial of degree $\leqq 2k - 1$, and so the derivative of arc tan $x - \overset{*}{P}_{2k}(x)$ cannot vanish more than $2k$ times; this implies that the points ± 1 are critical points, and, in addition, since this derivative must vanish $2k$ times, that $\overset{*}{P}_{2k}(x)$ is of degree $2k - 1$. Similarly, when $f(x) = \cos mx$, $m > 0$, $\overset{*}{P}_{2k+1}$ is an even polynomial of degree $2k$, the points 0 and 1 being critical points.

It is clear that $P_n{}^*(x)$ is easily determined if any set $x_1{}^* < x_2{}^* < \cdots < \overset{*}{x}_{n+2}$

Received May 14, 1959.

of critical points is known; indeed, the $n + 2$ equations $f(x_k^*) - P_n^*(x_k^*) = (-1)^{k-1}\omega$, $k = 1, \cdots, n + 2$, constitute a set of $n + 2$ linear equations for ω and the $n + 1$ coefficients of $P_n^*(x)$. If $x_1 < x_2 < \cdots < x_{n+2}$ is any set of $n + 2$ points of the interval $-1 \leqq x \leqq 1$, and we write $f(x_k) - P_n(x_k) = (-1)^{k-1}E$, $k = 1, \cdots, n + 2$, these equations determine E and the $n + 1$ coefficients of $P_n(x)$, and the function E of the $n + 2$ variables $(x_1, \cdots, x_{n+2})$ has an absolute maximum at $(x_1^*, \cdots, x_{n+2}^*)$. Thus, the derivative of E with respect to each of the n variables $x_2, \cdots, x_{n+1}$ is zero at $(x_1^*, \cdots, x_{n+2}^*)$, and this implies that the derivative of each of the $n + 1$ coefficients of $P_n(x)$ with respect to each of the n variables $x_2, \cdots, x_{n+1}$ is zero at $(x_1^*, \cdots, x_{n+2}^*)$. Hence, these coefficients are insensitive to small changes of the variables $(x_2, \cdots, x_{n+1})$ when these variables have the values $x_2^*, \cdots, x_{n+1}^*$ and $x_1 = x_1^*$, $x_{n+2} = x_{n+2}^*$.

The method by which we determine $P_n^*(x)$ is an iterative one. Let us suppose that the points ± 1 are critical points, so that there are n interior critical points, which we denote, changing slightly our previous notation, by $x_1^* < x_2^* < \cdots < x_n^*$. Let us suppose further, that we are in possession of a polynomial, $P_n^{(0)}(x)$, of degree $\leqq n$, which we term our entering polynomial and which possesses the following property: The difference $f(x) - P_n^{(0)}(x)$ assumes extreme values of alternating sign at $n + 2$ points of the interval $-1 \leqq x \leqq 1$. We denote by $x_1^{(1)}, \cdots, x_n^{(1)}$ approximations to the second, third, $\cdots$, $(n + 1)$st of these points and we regard $x_1^{(1)}, \cdots, x_n^{(1)}$ as approximations, in the first cycle of an iterative procedure, to $x_1^*, \cdots, x_n^*$. We determine the approximating polynomial of degree $\leqq n$, $P_n^{(1)}(x)$, with which we end the first, and begin the second, cycle of this procedure by means of the $n + 1$ linear equations obtained by eliminating $E^{(1)}$ from the $n + 2$ linear equations

$$f(-1) - P_n^{(1)}(-1) = E^{(1)}; \qquad f(x_k^{(1)}) - P_n^{(1)}(x_k^{(1)}) = (-1)^k E^{(1)},$$
$$k = 1, \cdots, n; \qquad f(1) - P_n^{(1)}(1) = (-1)^{n+1} E^{(1)}.$$

Denoting $f(-1) - P_n^{(0)}(-1)$ by $\delta_0^{(1)}$, $f(x_k^{(1)}) - P_n^{(0)}(x_k^{(1)})$ by $\delta_k^{(1)}$, $k = 1, \cdots, n$, and $f(1) - P_n^{(0)}(1)$ by $\delta_{n+1}^{(1)}$, we can write these $n + 2$ equations as

$$\delta P_n^{(0)}(-1) = \delta_0^{(1)} - E^{(1)}; \qquad \delta P_n(x_k^{(1)}) = \delta_k^{(1)} - (-1)^k E^{(1)},\ k = 1, \cdots, n;$$
$$\delta P_n^{(0)}(1) = \delta_{n+1}^{(1)} - (-1)^{n+1} E^{(1)},$$

where $\delta P_n^{(0)}(x)$ denotes the polynomial, of degree $\leqq n$, $P_n^{(1)}(x) - P_n^{(0)}(x)$. $E^{(1)}$ is conveniently eliminated by combining the last $n + 1$ of these $n + 2$ equations alternately by addition and subtraction with the first, and the $n + 1$ coefficients of $\delta P_n^{(0)}(x)$ are obtained by solving the resulting $n + 1$ linear equations. Then the coefficients of $P_n^{(1)}(x)$ are obtained by adding each of the coefficients of $\delta P_n^{(0)}(x)$ to the corresponding coefficients of $P_n^{(0)}(x)$.

The first step in the second cycle of the iterative procedure is the determination of new approximations $x_1^{(2)}, \cdots, x_n^{(2)}$ to $x_1^*, \cdots, x_n^*$. Just as $x_1^{(1)}, \cdots, x_n^{(1)}$ were approximations to the zeros of $D[f(x) - P_n^0(x)]$, where D denotes differentiation with respect to x, so $x_1^{(2)}, \cdots, x_n^{(2)}$ are approximations to the zeros of $D[f(x) - P_n^{(1)}(x)]$. Writing $x_k^{(2)} = x_k^{(1)} + \delta x_k^{(1)}$, $k = 1, \cdots, n$, we see that the value of $D[f(x) - P_n^{(0)}(x)]$ at $x_k^{(1)} + \delta x_k^{(1)}$ must be the same as the value of $D[\delta P_n^0(x)]$ at $x_k^{(1)} + \delta x_k^{(1)}$, and this is the same, to the first order of infinitesimals, as the value of $D[\delta P_n^{(0)}(x)]$ at $x_k^{(1)}$. Thus, to the first order of infinitesimals, the value of $D[f(x) - P_n^{(0)}(x)]$ at $x_k^{(1)}$ plus the value of $D^2[f(x) - P_n^{(0)}(x)]$ at $x_k^{(1)}$ times $\delta x_k^{(1)}$ is equal to the value of $D[\delta P_n^{(0)}(x)]$ at $x_k^{(1)}$, so that $\delta x_k^{(1)}$ is the negative of the quotient of the value of $D[f(x) - P_n^{(1)}(x)]$ at $x_k^{(1)}$ by the value of $D^2[f(x) - P_n^{(0)}(x)]$ at $x_k^{(1)}$. We denote the value of $D[f(x) - P_n^{(1)}(x)]$ at $x_k^{(1)}$ by $\epsilon_k^{(2)}$, $k = 1, \cdots, n$, so that $\delta x_k^{(1)}$ is the negative of the quotient of $\epsilon_k^{(2)}$ by the value of $D^2[f(x) - P_n^{(0)}(x)]$ at $x_k^{(1)}$. If $\epsilon_k^{(2)}$ is zero, $x_k^{(2)} = x_k^{(1)}$ and $D[f(x) - P_n^{(1)}(x)]$ is zero at $x_k^{(2)} = x_k^{(1)}$. In order to complete the second cycle, we calculate the $n + 2$ numbers $\delta_0^{(2)} = f(-1) - P_n^{(1)}(-1)$, $\delta_k^{(2)} = f(x_k^{(2)}) - P_n^{(1)}(x_k^{(2)})$, $(k = 1, \cdots, n)$, $\delta_{n+1}^{(2)} = f(1) - P_n^{(1)}(1)$, and determine, as before, the coefficients of $\delta P_n^{(1)}(x)$. If, to the number of decimals we are using, the $n + 2$ numbers $\delta_0, \cdots, \delta_{n+1}$ are equal in absolute value and alternating in sign, the coefficients of $\delta P_n^{(1)}(x)$ are all zero and the approximating polynomial, $P_n^{(2)}(x)$, with which we end the second cycle is the same as the approximating polynomial, $P_n^{(1)}(x)$, with which we began it.

In a previous publication [2] we determined the numbers $x_1, \cdots, x_n$ in each cycle by solving the equation $D[f(x) - P_n(x)] = 0$, where $P_n(x)$ is the approximating polynomial, of degree $\leqq n$, with which we begin the cycle, but the less exacting method of the present paper is equally effective.

It remains only to describe the selection of our entering polynomial approximation, $P_n^{(0)}(x)$, of degree $\leqq n$, and the determination of the approximations $x_1^{(1)}, \cdots, x_n^{(1)}$ to the n points of the interval $-1 < x < 1$ at which $D[f(x) - P_n^{(0)}(x)]$ is zero. On setting $x = \cos \theta$, we see that $f(x)$ becomes a function, $F(\theta)$, of θ defined over $0 \leqq \theta \leqq \pi$, and we write the Fourier cosine series of $F(\theta)$ as $\frac{1}{2}a_0 + a_1 \cos \theta + a_2 \cos 2\theta + \cdots$. Then $\cos m\theta$ is a polynomial function, $T_m(x)$, of x of degree m, which is known as the mth Chebyshev polynomial, $m = 0, 1, 2, \cdots$, and $\frac{1}{2}a_0 + a_1T_1(x) + a_2T_2(x) + \cdots$ is known as the Chebyshev expansion of $f(x)$. The sum of the first $n + 1$ terms of this Chebyshev expansion of $f(x)$ is a polynomial function, of degree $\leqq n$, of x, and it is this polynomial function that we take as $P_n^{(0)}(x)$. We say that $P_n^{(0)}(x)$ is furnished by the truncated Chebyshev expansion (the truncation taking place at the term which involves $T_n(x)$). Now $f(x) - P_n^{(0)}(x) = a_{n+1}T_{n+1}(x) + \cdots$, and we take as our approximations to the n points of the interval $-1 < x < 1$ at which $D[f(x) - P_n^{(0)}(x)] = 0$ the n points of this interval at which $D[T_{n+1}(x)] = 0$, it being assumed that $a_{n+1} \neq 0$. (If $f(x)$ is odd its Chebyshev expansion is of the form $a_1T_1(x) + a_3T_3(x) + \cdots$ and $n = 2m$ is even; then we truncate this Chebyshev expansion at the term involving $T_{2m-1}(x)$, and we take as our approximations to the m points of the interval $0 < x < 1$ at which $D[f(x) - P_n^{(0)}(x)]$ is zero the m points of this interval at which $D[T_{2m+1}(x)]$ is zero. Similar remarks apply to the case where $f(x)$ is even and $n = 2m + 1$ is odd.) Since

$$D\,[T_{n+1}(x)] = (n + 1)\,\frac{\sin\,(n + 1)\,\theta}{\sin \theta},$$

we have

$$x_k^{(1)} = \cos\left(\pi - \frac{k\pi}{n + 1}\right), \qquad k = 1, \cdots, n.$$

Example 1. $f(x) = \text{arc} \tan x$, $n = 6$.

There are three positive interior critical points, which we denote by x_1^*, x_2^*, x_3^*, and

$$P_6^{(0)}(x) = 0.994949366x - 0.287060636x^3 + 0.078937176x^5,$$

since the Chebyshev expansion of arc tan x is

$$2\left[pT_1(x) - \frac{p^3}{3}T_3(x) + \frac{p^5}{5}T_5(x) - \cdots\right],$$

where $p = 2^{\frac{1}{2}} - 1 = 0.414213562$, to 9 decimals. The first-cycle approximations to x_1^*, x_2^*, x_3^* are

$$x_1^{(1)} = 0.222520934, \qquad x_2^{(1)} = 0.623489802, \qquad x_3^{(1)} = 0.900968868,$$

and the polynomial approximation with which we end the first cycle is

$$P_6^{(1)}(x) = 0.995383022x - 0.288700440x^3 + 0.079313307x^5,$$

the values of arc tan $x - P_n^{(0)}(x)$ at the points $x_1^{(1)}$, $x_2^{(1)}$, $x_3^{(1)}$, 1 being

$$\delta_1^{(1)} = 0.000676851, \qquad \delta_2^{(1)} = -0.000604555,$$
$$\delta_3^{(1)} = 0.000546760, \qquad \delta_4^{(1)} = -0.000527744,$$

respectively. The corresponding results for the second cycle are

$$x_1^{(2)} = 0.205422893, \qquad x_2^{(2)} = 0.593832571, \qquad x_3^{(2)} = 0.888813502,$$
$$\delta_1^{(2)} = 0.000603543, \qquad \delta_2^{(2)} = -0.000619441,$$
$$\delta_3^{(2)} = 0.000607728, \qquad \delta_4^{(2)} = -0.000597725,$$

and

$$P_6^{(2)}(x) = 0.995357994x - 0.288690417x^3 + 0.079339173x^5.$$

In the third cycle we find

$$x_1^{(3)} = 0.205218790, \qquad x_2^{(3)} = 0.593469973, \qquad x_3^{(3)} = 0.888196372,$$
$$\delta_1^{(3)} = 0.000608588, \qquad \delta_2^{(3)} = -0.000608590,$$
$$\delta_3^{(3)} = 0.000608612, \qquad \delta_4^{(3)} = -0.000608588,$$

and

$$P_6^{(3)}(x) = 0.995357955x - 0.288690238x^3 + 0.079339041x^5.$$

Finally, in the fourth cycle we obtain

$$x_1^{(4)} = 0.205219373, \qquad x_2^{(4)} = 0.593470162, \qquad x_3^{(4)} = 0.888196289,$$
$$\delta_1^{(4)} = 0.000608595, \qquad \delta_2^{(4)} = -0.000608595, \qquad \delta_3^{(4)} = 0.000608595,$$
$$\delta_4^{(4)} = -0.000608595, \qquad P_6^{(4)}(x) = P_6^{(3)}(x).$$

Thus, to seven decimals,

$$P_6^*(x) = 0.9953580x - 0.2886902x^3 + 0.0793390x^5,$$

and the maximum of $|\arctan x - P_6^*(x)|$ over $-1 \leqq x \leqq 1$ is 0.0006086. Hastings [3] has given $P_6^*(x)$ for arc tan x to six decimals as $0.995354x - 0.288679x^3 + 0.079331x^5$.

Example 2. $f(x) = \log \dfrac{a+x}{a-x}, \qquad a = \dfrac{10^{\frac{1}{2}}+1}{10^{\frac{1}{2}}-1}, \qquad n = 4.$

The number a must be greater than 1; we use the value indicated in order to check the work of Hastings. Setting $\xi = (a+x)/(a-x)$, the polynomial $P_4^*(x)$ which we determine will be an approximation to $\log \xi$ over the interval $10^{-\frac{1}{2}} \leqq \xi \leqq 10^{\frac{1}{2}}$.

The Chebyshev expansion of $\log (a+x)/(a-x)$ is

$$4M\left[pT_1(x) + \frac{p^3}{3}T_3(x) + \frac{p^5}{5}T_5(x) + \cdots\right],$$

where $M = \log e = 0.434294482$, to 9 decimals, and $p = a - (a^2-1)^{\frac{1}{2}} = 0.280130000$. The entering polynomial approximation is $P_4^{(0)}(x) = 0.448447982x + 0.050916894x^3$, and our first-cycle approximations to the two positive interior critical points are

$$x_1^{(1)} = 0.309016994; \qquad x_2^{(1)} = 0.809016994.$$

The values of $\log (a+x)/(a-x) - P_4^{(0)}(x)$ at the points $x_1^{(1)}$, $x_2^{(1)}$, 1 are

$$\delta_1^{(1)} = 0.000572828, \qquad \delta_2^{(1)} = -0.000607958, \qquad \delta_3^{(1)} = 0.000635124,$$

respectively, and $P_4^{(1)}(x) = 0.448349355x + 0.051051305x^3$. In the second cycle

$$x_1^{(2)} = 0.321484228; \qquad x_2^{(2)} = 0.821954759$$

$$\delta_1^{(1)} = 0.000600487, \qquad \delta_2^{(2)} = -0.000602901, \qquad \delta_3^{(2)} = 0.000599339$$

$$P_4^{(2)}(x) = 0.448347007x + 0.051051766x^3,$$

and, in the third cycle,

$$x_1^{(3)} = 0.321320097; \qquad x_2^{(3)} = 0.821455202$$

$$\delta_1^{(3)} = 0.000601227; \qquad \delta_2^{(3)} = -0.000601233; \qquad \delta_3^{(3)} = 0.000601227$$

$$P_4^{(3)}(x) = 0.448346999x + 0.051051771x^3,$$

so that, to seven decimals, $P_4^{(3)}(x)$ coincides with $P_4^{(2)}(x)$, and $P_4^*(x) = 0.4483470x + 0.0510518x^3$, the maximum value of $|\log (a+x)/(a-x) - P_4^*(x)|$ over $-1 \leqq x \leqq 1$ being 0.0006012.

If x is replaced by $a(x-1)/(x+1)$, there results the approximation 0.8630458 $[(x-1)/(x+1)] + 0.3641410[(x-1)/(x+1)]^3$ to $\log x$ over the interval $10^{-\frac{1}{2}} \leqq x \leqq 10^{\frac{1}{2}}$. Hastings [3] gives as the coefficients in this approximation the numbers 0.86304 and 0.36415, respectively.

In a recent paper by Barth [4], $P_6^*(x)$ for $\ln (a+x)/(a-x)$, $a = (10^{\frac{1}{2}}+1)/(10^{\frac{1}{2}}-1)$, is given, to ten decimals, as

$$0.8690286986x + 0.2773833195x^3 + 0.2543282307x^5.$$

The correct formula, to seven decimals, for $P_6^*(x)$ is

$$0.8690285x + 0.2773864x^3 + 0.2543195x^5,$$

the maximum of $|\ln (a + x)/(a - x) - P_6^*(x)|$ over $-1 \leqq x \leqq 1$ being 0.0000337.

Example 3. $f(\xi) = \ln (1 + \xi),\ 0 \leqq \xi \leqq 1,\ n = 4.$

We denote in this example the independent variable by ξ, instead of x, since the interval, $0 \leqq \xi \leqq 1$, is not the standard interval $-1 \leqq x \leqq 1$. The linear transformation $x = 2\xi - 1$ transforms the interval $0 \leqq \xi \leqq 1$ into the interval $-1 \leqq x \leqq 1$, and the problem of determining $P_4^*(\xi)$ for $\ln (1 + \xi)$ over $0 \leqq \xi \leqq 1$ is the same as that of determining $P_4^*(x)$ for $\ln \frac{1}{2}(3 + x)$ over $-1 \leqq x \leqq 1$.

The Chebyshev expansion of $\ln \frac{1}{2}(3 + x)$ is

$$-2 \log_e 2p + 2\left[p^2T_1(x) - \frac{p^4}{2}\, T_2(x) + \frac{p^6}{3}\, T_3(x) \cdots\right],$$

where $p = 2^{\frac{1}{2}} - 1$, and our entering polynomial approximation, of degree 4, is $P_4^{(0)}(\xi) = 0.000069446 + 0.996261948\xi - 0.466442439\xi^2 + 0.218665484\xi^3 - 0.055459314\xi^4$. Our first-cycle approximations to the four interior critical points are

$$\xi_1^{(1)} = 0.095491503, \qquad \xi_2^{(1)} = 0.345491503,$$
$$\xi_3^{(1)} = 0.654508497, \qquad \xi_4^{(1)} = 0.904508497,$$

and the values of $\ln (1 + \xi) - P_4^{(0)}(\xi)$ at the points $0, \xi_1^{(1)}, \xi_2^{(1)}, \xi_3^{(1)}, \xi_4^{(1)}, 1$ are

$$\delta_0^{(1)} = -0.000069446, \qquad \delta_1^{(1)} = 0.000066650, \qquad \delta_2^{(1)} = -0.000060948,$$
$$\delta_3^{(1)} = 0.000055988, \qquad \delta_4^{(1)} = -0.000053017, \qquad \delta_5^{(1)} = 0.000052055.$$

The polynomial approximation, of the fourth degree, with which we end the first cycle is, to eight decimals,

$$P_4^{(1)}(\xi) = 0.000059471 + 0.996558114\xi - 0.467864445\xi^2$$
$$+ 0.220882267\xi^3 - 0.056547698\xi^4.$$

In the second cycle we obtain

$$\xi_1^{(2)} = 0.084407707, \qquad \xi_2^{(2)} = 0.318071278,$$
$$\xi_3^{(2)} = 0.629216597, \qquad \xi_4^{(2)} = 0.895475308;$$
$$\delta_0^{(2)} = -0.000059471, \qquad \delta_1^{(2)} = 0.000060703, \qquad \delta_2^{(2)} = -0.000061938;$$
$$\delta_3^{(2)} = 0.000061315, \qquad \delta_4^{(2)} = -0.000060043, \qquad \delta_5^{(2)} = 0.000059471;$$
$$P_4^{(2)}(\xi) = 0.000060712 + 0.996540728\xi - 0.467834593\xi^2$$
$$+ 0.220891205\xi^3 - 0.056571583\xi^4;$$

and, in the third cycle,

$$\xi_1^{(3)} = 0.085058286, \qquad \xi_2^{(3)} = 0.319106141, \qquad \xi_3^{(3)} = 0.629174645,$$
$$\xi_4^{(3)} = 0.895122761;$$

$$\delta_0^{(3)} = -0.000060712, \quad \delta_1^{(3)} = 0.000060716, \quad \delta_2^{(3)} = -0.000060716,$$
$$\delta_3^{(3)} = 0.000060712, \quad \delta_4^{(3)} = -0.000060713, \quad \delta_5^{(3)} = 0.000060712;$$
$$P_4^{(3)}(\xi) = 0.000060714 + 0.996540741\xi - 0.467834762\xi^2 + 0.220891541\xi^3 - 0.056571768\xi^4.$$

Beginning the fourth cycle, we compute

$$\xi_1^{(4)} = 0.085060350, \quad \xi_2^{(4)} = 0.319112305,$$
$$\xi_3^{(4)} = 0.629171981, \quad \xi_4^{(4)} = 0.895123131,$$

and we find that the corresponding values $\delta_i^{(4)}$ for $i = 0, 1, 2, 3$, and 4 are all numerically equal to 0.0000607141, to within a unit in the tenth decimal place.

Thus, to seven-decimal accuracy in the coefficients, we have

$$P_4^*(\xi) = 0.0000607 + 0.9965407\xi - 0.4678348\xi^2 + 0.2208915\xi^3 - 0.0565718\xi^4.$$

The discrepancy between this approximation and the similar one presented in our earlier paper [2] is attributable to the premature termination of the iterative procedure in that reference, which stemmed from the erroneous belief that the precision of the coefficients of the approximating polynomial was comparable to that of the maximum difference between that polynomial and the given function. In this example the quantities $\delta_i^{(4)}$ have all become stabilized to 10 decimal places, whereas the coefficients of the corresponding approximating polynomial are subject to errors of approximately a unit in the eighth decimal place. This behavior of the coefficients is due to the relatively small value of the determinant of the system of equations used for their evaluation. Calculation of such coefficients to ten-place accuracy generally will require double-precision operations.

Hastings [3] gives as an approximating polynomial of degree 4, whose graph is arbitrarily required to pass through the origin, the following

$$0.9974442x - 0.4712839x^2 + 0.2256685x^3 - 0.0587527x^4,$$

for which the maximum departure from $\ln(1 + x)$ over the interval $0 \leqq x \leqq 1$ is 0.0000710, in contrast to the value 0.0000607, attained by the Chebyshev approximating polynomial of the same degree.

Example 4. $f(x) = \cos(\pi/4)x$, $n = 3$.

The Chebyshev expansion of $\cos(\pi/4)x$ is $J_0(\pi/4) - 2J_2(\pi/4)T_2(x) + 2J_4(\pi/4)T_4(x) - \cdots$, where $J_{2k}(\pi/4)$, for $k = 0, 1, \cdots$, is the value at $\pi/4$ of the Bessel function of the first kind, of order $2k$. There is only one positive interior critical point x^*, and our first-cycle approximation to this is $x^{(1)} = \cos \pi/4 = 0.707106781$. Our entering polynomial approximation is $P_3^{(0)}(x) = 0.998068558 - 0.292873289x^2$. The values of $\cos(\pi/4)x - P_3^{(0)}(x)$ at the points 0, $x^{(1)}$, 1 are

$$\delta_0^{(1)} = 0.001931442, \quad \delta_1^{(1)} = -0.001921422, \quad \delta_2^{(1)} = 0.001911512,$$

and $P_3^{(1)}(x) = 0.998078551 - 0.292893219x^2$. The coefficient of x^2 remains the same in all the succeeding cycles, so that, in this example, we obtain the coefficient of x^2 in $P_3^*(x)$ before we begin the second cycle; this simplification is due to the fact that both the points 0 and 1 are critical points, and this implies that the coefficient of x^2 in $P_3^*(x)$ is $\cos \pi/4 - 1$. In the second cycle we obtain

$$x^{(2)} = 0.705276652, \quad \delta_0^{(2)} = 0.001921449, \quad \delta_1^{(2)} = -0.001921553,$$

$$\delta_2^{(2)} = \delta_0^{(2)}, \quad P_3^{(2)}(x) = 0.998078499 - 0.292893219x^2;$$

and, in the third cycle,

$$x^{(3)} = 0.705270859, \quad \delta_0^{(3)} = 0.001921501, \quad \delta_1^{(3)} = -0.001921500,$$

$$\delta_2^{(3)} = \delta_0^{(3)}, \quad P_3^{(3)}(x) = 0.998078499 - 0.292893219x^2.$$

Thus, to seven decimals, $P_3^*(x) = 0.9980785 - 0.2928932x^2$, the maximum of $| \cos (\pi/4)x - P_3^*(x) |$ over $-1 \leqq x \leqq 1$ being 0.0019215.

We observe in this example that the entering polynomial approximation, $P_3^{(0)}(x)$, is so good that the maximum of $| \cos (\pi/4)x - P_3^{(0)}(x) |$, over $-1 \leqq x \leqq 1$, is 0.0019314, which exceeds the corresponding maximum of $| \cos (\pi/4)x - P_3^*(x) |$ by less than 0.52 per cent.

Example 5. $f(x) = \cos (\pi/2)x$, $n = 5$.

There are two positive interior critical points, x_1^* and x_2^*, and our first-cycle approximations are $x_1^{(1)} = \cos \pi/3 = \frac{1}{2}$ and $x_2^{(1)} = \cos \pi/6 = 3^{\frac{1}{2}}/2$. The Chebyshev expansion of $\cos (\pi/2)x$ is $J_0(\pi/2) - 2J_2(\pi/2)T_2(x) + \cdots$, and our entering polynomial approximation, of degree four, is $P_5^{(0)}(x) = J_0(\pi/2) - 2J_2(\pi/2)T_2(x) + 2J_2(\pi/4)T_4(x) = 0.999396554 - 1.222743153x^2 + 0.223936637x^4$. The values of $\cos (\pi/2)x - P_5^{(0)}(x)$ at the points 0, $x_1^{(1)}$, $x_2^{(1)}$, 1 are

$$\delta_0^{(1)} = 0.000603446, \quad \delta_1^{(1)} = -0.000600024,$$

$$\delta_2^{(1)} = 0.000593320, \quad \delta_3^{(1)} = -0.000590037.$$

Since 0 and 1 are critical points, the coefficients of the approximating polynomial $P_5^{(k)}(x)$, with which we end the kth cycle ($k = 1, 2, \cdots$) satisfy the relation $2\alpha^{(k)} + \beta^{(k)} + \gamma^{(k)} = 1$, and this implies that the coefficients of $P_5^*(x)$ satisfy the relation $2\alpha^* + \beta^* + \gamma^* = 1$. We find that

$$P_5^{(1)}(x) = 0.999403304 - 1.22796880x^2 + 0.223990272x^4.$$

In the second cycle we obtain

$$x_1^{(2)} = 0.497202761, \quad x_2^{(2)} = 0.864404535;$$

$$\delta_0^{(2)} = 0.000596695, \quad \delta_1^{(2)} = -0.000596808,$$

$$\delta_2^{(2)} = 0.000596805, \quad \delta_3^{(2)} = -0.000596697;$$

$$P_5^{(2)}(x) = 0.999403231 - 1.222796733x^2 + 0.223990272x^4.$$

In the third cycle we obtain

$$x_1^{(3)} = 0.497195260, \quad x_2^{(3)} = 0.864395279;$$

$$\delta_0^{(3)} = 0.000596769, \quad \delta_1^{(3)} = -0.000596772,$$

$$\delta_2^{(3)} = 0.000596770, \quad \delta_3^{(3)} = -0.000596770;$$

$$P_5^{(3)}(x) = P_5^{(2)}(x) \text{ to 9 decimals.}$$

Hence, we conclude that

$$P_5^*(x) = 0.9994032 - 1.2227967x^2 + 0.2239903x^4,$$

the maximum of $| \cos (\pi/2)x - P_5^*(x) |$ over $-1 \leqq x \leqq 1$ being 0.0005968.

The iterative procedure described herein for the determination of the Chebyshev approximating polynomial $P_n^*(x)$, of degree not exceeding n, over the interval $-1 \leqq x \leqq 1$, to a given differentiable function $f(x)$ will converge if the difference between $f(x)$ and the initial approximating polynomial $P_n^{(0)}(x)$ assumes extreme values at $n + 2$ points of the interval $-1 \leqq x \leqq 1$ and if, furthermore, these extreme values alternate in sign. A proof of this based on the argument of Novodvorskii and Pinsker [5] has been given, and illustrated by a numerical example, in our previous publication [2] on this subject.

Applied Mathematics Laboratory,
David Taylor Model Basin,
Washington 7, District of Columbia

1. C. de la Vallée Poussin, *Leçons sur l'Approximation des Fonctions d'une Variable Réelle*, Gauthier-Villars, Paris, 1952.

2. F. D. Murnaghan & J. W. Wrench, Jr., *The Approximation of Differentiable Functions by Polynomials*, David Taylor Model Basin Report 1175, April 1958.

3. Cecil Hastings, Jr., *Approximations for Digital Computers*, Princeton University Press, Princeton, New Jersey, 1955.

4. W. Barth, "Ein iterationsverfahren zur approximation durch polynome," *Zeitschrift für Angewandte Mathematik und Mechanik*, v. 38, 1958, p. 258–260.

5. E. P. Novodvorskii & I. Sh. Pinsker, "On a process of equalization of maxima" (in Russian), *Uspekhi Matematicheskikh Nauk*, v. 6, 1951, p. 174–181.

A Method of Computing Eigenvectors and Eigenvalues on an Analog Computer

By Lucien Neustadt*

1. Introduction. Many papers have been written in recent years describing methods for finding the eigenvalues and eigenvectors of an arbitrary matrix. Most of these methods apply to digital computation, and little attention has been paid to methods applicable to analog computers. One such method, which has been used successfully in the past, is described in [2]. The analog technique described in the present paper has the advantage of yielding both eigenvectors and eigenvalues of a real symmetric, or complex Hermitian matrix, and of not requiring a trial and error procedure. This is accomplished at the expense of additional computing equipment.

2. Mathematical Formulation. Consider the real, symmetric, $\mathbf{n} \times \mathbf{n}$ matrix $A = (a_{ij})$ with $a_{ij} = a_{ji}$. We shall state, without proof, the following properties of the matrix A [3], [4]:

(1) There is an orthonormal set of n real eigenvectors $\mathbf{e}^1, \mathbf{e}^2, \cdots, \mathbf{e}^n$ of the matrix A; i.e., there exist n numbers $\lambda_1, \lambda_2, \cdots, \lambda_n$, (the eigenvalues) with $A\mathbf{e}^i = \lambda_i\mathbf{e}^i$ and $i = 1, 2, \cdots, n$; also

$$\mathbf{e}^i \cdot \mathbf{e}^j = \delta_i^{\ j}.$$

(2) The eigenvalues are all real.

Suppose that the eigenvalues are arranged in descending order such that $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n$; then for an arbitrary real, non-zero vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}:$$

$$\lambda_1 = \sup_{\mathbf{x}} \frac{\sum_{i,j=1}^{n} a_{ij}\, x_i\, x_j}{\sum_{i=1}^{n} x_i^2} \equiv \sup_{\mathbf{x}} \frac{A\mathbf{x}\cdot\mathbf{x}}{\mathbf{x}\cdot\mathbf{x}} = \sup_{\mathbf{x}\cdot\mathbf{x}=1} A\mathbf{x}\cdot\mathbf{x} \tag{3}$$

The sup is attained at an eigenvector $\mathbf{e}^1$, belonging to $\lambda_1 \cdot \mathbf{e}^i$ and λ_i for $i > 1$ can be obtained from the same formula if the following added restriction is made:

$$\mathbf{x}\cdot\mathbf{e}^j = 0 \qquad j = 1, 2, \cdots, i - 1.$$

$\mathbf{e}^n$ and λ_n can also be computed from formula (3) if the sup is replaced by inf.

Received March 7, 1958, in revised form, August 26, 1958.

* Now at Space Technology Laboratories, Los Angeles, California.

3. Solution Procedure. The technique is based on the analog computer method for solving problems in linear programming [5]. This method may be readily adapted to problems in nonlinear programming, i.e. to the problem of finding the extreme values of a nonlinear function subject to nonlinear restrictions. The problem of finding eigenvalues and eigenvectors is just such a problem, namely that of locating the extreme values listed above.

Let $\mathbf{x}$ be the radius vector to the point $x = (x_1, \cdots, x_n)$ in n space. Let x_1, $x_2, \cdots, x_n$ be functions of time. As described in the mathematical formulation, $\mathbf{e}^1$ may be found as follows:

Let the point x move on the unit sphere until it reaches a steady state corresponding to a maximum of the function $A\mathbf{x}\cdot\mathbf{x}$. Then $\mathbf{e}^1$ equals the steady state of $\mathbf{x}$.

The vector $\mathbf{e}^k$ is found in the same way, with the additional restrictions on $\mathbf{x}$ that $\mathbf{x}\cdot\mathbf{e}^1 = \mathbf{x}\cdot\mathbf{e}^2 = \cdots = \mathbf{x}\cdot\mathbf{e}^{k-1} = 0$.

Also, $\mathbf{e}^n$ is found in the same manner as $\mathbf{e}^1$, except that the steady state now corresponds to a minimum of $A\mathbf{x}\cdot\mathbf{x}$.

In order to find $\mathbf{e}^1$, we write a differential equation for the vector $\mathbf{x}$ such that the steady state solution of the equation is $\mathbf{e}^1$.

Denote $A\mathbf{x}\cdot\mathbf{x}$ by $f(x) = f(x_1, x_2, \cdots, x_n)$.

If we set $\dot{\mathbf{x}} = \operatorname{grad} f$, the point x will move in such a way as to increase f. To insure a steady state with $\mathbf{x}\cdot\mathbf{x} = 1$, modify this equation to read

$$\dot{\mathbf{x}} = \operatorname{grad} f + k\epsilon\mathbf{x} \tag{1}$$

where

$$\epsilon = \begin{cases} 1 \text{ if } \mathbf{x}\cdot\mathbf{x} \leqq 1 \\ -1 \text{ if } \mathbf{x}\cdot\mathbf{x} > 1 \end{cases}$$

and k is a positive constant, chosen such that

$$k > 2 \max (|\lambda_1|, |\lambda_n|).$$

The point x will move as follows: Assume $x(0) = 0$. The origin is a point of unstable equilibrium. Once x moves slightly away from zero it will continue to move in a direction which is given by the vector sum of grad f and $k\mathbf{x}$. Because k is large enough, x will move toward the boundary of the unit sphere. Once it breaks through, again because of the choice of k, it will immediately re-enter. Having reentered, it will immediately break through again, and so on indefinitely. The point x will then oscillate back and forth across the surface of the sphere with a frequency depending on the rapidity of the switching arrangement which determines the sign of ϵ. Superimposed on this oscillation is a tangential motion produced by the tangential component of grad f. The tangential motion will continue until a point is reached where grad f has only a normal component. There grad $f(x) = a\mathbf{x}$ for some constant a. It is easily verified that in general, grad $f(x) = 2A\mathbf{x}$, so that this point corresponds to an eigenvector. Furthermore, at such a point, $f(x)$, where x is restricted to lie on the unit sphere, has an extreme value. If $\mathbf{x} \neq \mathbf{e}^1$, the solution is unstable, since any motion of x in the direction of $\mathbf{e}^1$ will be self-reinforcing. A stable steady state solution is reached only when $\mathbf{x} = \mathbf{e}^1$.

The eigenvector $\mathbf{e}^n$, belonging to the smallest eigenvalue λ_n, is the steady state of the solution to the equation

$$\dot{\mathbf{x}} = -\operatorname{grad} f + k\epsilon\mathbf{x}. \tag{2}$$

If equation (1) is modified, the steady state of $\mathbf{x}$ can be made to correspond to the eigenvector $\mathbf{e}^2$. Consider the differential equation:

$$\dot{\mathbf{x}} = \operatorname{grad} f + k\epsilon\mathbf{x} + k_1\epsilon_1 \operatorname{grad} f_1 ; \tag{3}$$

where (a) ϵ and k are defined as above.

(b) $$f_1(x) = \mathbf{x}\cdot\mathbf{e}^1,$$

(c) $$\epsilon_1 = \begin{cases} 1 & \text{if } f_1(x) \leqq 0 \\ -1 & \text{if } f_1(x) > 0, \quad \text{and} \end{cases}$$

(d) k_1 is a large enough positive constant. It is sufficient that $k_1 > 2k$.

The added term will insure that the steady state of x yield a radius vector $\mathbf{x}$ orthogonal to $\mathbf{e}^1$. Since the maximum condition is also satisfied, this steady state corresponds to the eigenvector $\mathbf{e}^2$.

The other eigenvectors can be obtained by modifying equations 2 or 3 by the addition of similar terms.

In principle the eigenvalues can be obtained directly from the eigenvectors. Given the eigenvector $\mathbf{e}^i$, one may compute $A\mathbf{e}^i$. The ratio of the components of $A\mathbf{e}^i$ to the components of $\mathbf{e}^i$ equals the eigenvalue λ_i. Since $\mathbf{e}^i$ is obtained only approximately, this ratio will in general not be constant.

An excellent approximation to the eigenvalue, knowing an approximate eigenvector $\mathbf{e}^i$, is given by

$$\frac{A\mathbf{e}^i\cdot\mathbf{e}^i}{\mathbf{e}^i\cdot\mathbf{e}^i}$$ [4].

4. The Computer Setup. In order to find the eigenvector corresponding to the largest (or smallest) eigenvalue of A the following amount of computing equipment is necessary:

n integrating amplifiers, whose outputs are $x_1, x_2, \cdots, x_n$.

n inverting amplifiers, whose outputs are $-x_1, -x_2, \cdots, -x_n$.

n^2 scale factor potentiometers, corresponding to n potentiometers for each component of grad f.

One "switch" to compute ϵ. A high gain amplifier with two diodes and two voltage sources can be used for the switch.

Multiplying equipment to compute $x_1^2, x_2^2, \cdots, x_n^2$, as well as $\epsilon x_1, \epsilon x_2, \cdots, \epsilon x_n$. For this purpose n multiplying servos positioned by $x_1, \cdots, x_n$ may be used.

One inverting amplifier to compute $-\epsilon$.

In order to find the eigenvector corresponding to λ_2 (or λ_{n-1}) the following additional equipment is necessary:

A switch to compute ϵ_1. This can again be a high gain amplifier.

n potentiometers to compute $f_1(x)$.

n potentiometers to compute the components of grad f_1.

One inverter to compute $-\epsilon_1$.

For every additional pair of eigenvectors more of the same equipment must be used.

The eigenvalues can also be found with the computer. Having computed the approximate unit eigenvector $\mathbf{e}^i$, λ_i can be approximated by $A\mathbf{e}^i \cdot \mathbf{e}^i$. But the steady state of $\mathbf{x}$ is $\mathbf{e}^i$, and $A\mathbf{x} = \frac{1}{2}\,\mathrm{grad}\, f(x)$, so that in the steady state $2\lambda_i = (\mathrm{grad}\, f) \cdot \mathbf{x}$. Since the components of grad f are available from the computation of $\dot{\mathbf{x}}$, λ_i can be obtained by performing the n multiplications necessary to form the dot product. The amount of equipment necessary for this is:

n summing amplifiers, whose outputs are the components of grad f.

n inverting amplifiers, whose outputs are the negative of the components of grad f.

One summing amplifier to form the dot product. Multiplying equipment to compute the terms of the dot product $\mathbf{x} \cdot (\mathrm{grad}\, f)$. If multiplying servos were used above, one additional potentiometer per shaft is required.

Two precautions must be taken in order to allow for the dynamic limitations of the multipliers. First, the problem may have to be slowed down in order that the multipliers remain within their rate limits. Second, the constant k may have to be adjusted carefully, in order to insure stability of the computing loop.

Once an $n \times n$ matrix problem has been programmed, this programming—and the associated patching—can be used for any other $n \times n$ matrix, with only minor changes. Since the x_i are restrained to lie between 1 and -1, the scaling need not be changed. Only the potentiometer settings and the sign of the variable being fed into the potentiometers must be adjusted.

As an illustration, consider the following 3×3 matrix:

$$A = \begin{bmatrix} 3 & 5 & 2 \\ 5 & 0 & -3 \\ 2 & -3 & 1 \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$f(x) = A\mathbf{x} \cdot \mathbf{x} = 3x_1^2 + 10x_1x_2 + 4x_1x_3 - 6x_2x_3 + x_3^2$$

$$\mathrm{grad}\, f = \begin{bmatrix} 6x_1 & + & 10x_2 & + & 4x_3 \\ 10x_1 & - & 6x_3 & & \\ 4x_1 & - & 6x_2 & + & 2x_3 \end{bmatrix}$$

Let $k = 20$.

Equation (1) becomes

$$\begin{cases} \dot{x}_1 = 6x_1 + 10x_2 + 4x_3 + 20\epsilon x_1 \\ \dot{x}_2 = 10x_1 - 6x_3 + 20\epsilon x_2 \\ \dot{x}_3 = 4x_1 - 6x_2 + 2x_3 + 20\epsilon x_3 , \end{cases}$$

where

$$\epsilon = \mathrm{sgn}\, [1 - (x_1^2 + x_2^2 + x_3^2)].$$

The programming for this example is shown in Figure 1. The problem has been slowed down by a factor of twenty.

The steady state of this system yields the eigenvector $\mathbf{e}^1$ corresponding to the largest eigenvalue λ_1.

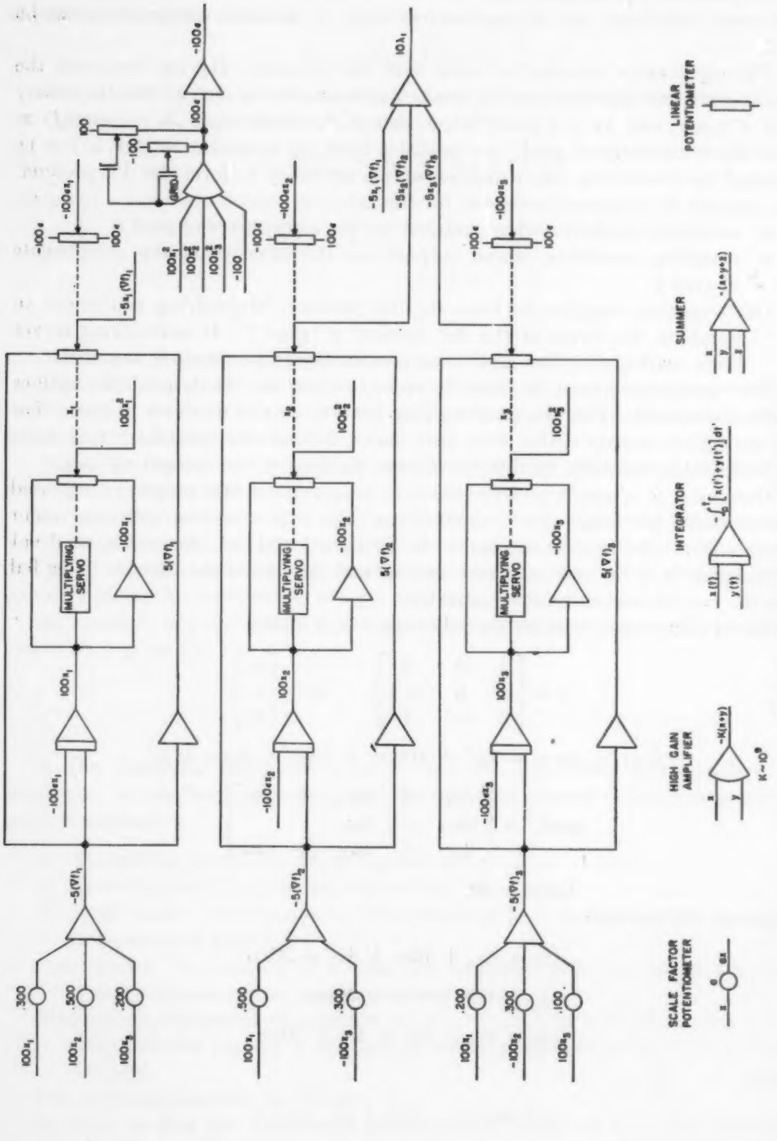

FIG. 1.—Program for typical problem.

In order to compute the eigenvector $\mathbf{e}^3$ corresponding to the smallest eigenvalue λ_3, equation 3 is used:

$$\begin{cases} \dot{x}_1 = -6x_1 - 10x_2 - 4x_3 + 20\epsilon x_1 \\ \dot{x}_2 = -10x_1 + 6x_3 + 20\epsilon x_2 \\ \dot{x}_3 = -4x_1 + 6x_2 - 2x_3 + 20\epsilon x_3 . \end{cases}$$

Finally, in order to compute the eigenvector $\mathbf{e}^2$ either of the two above systems of equations must be modified by the addition of an extra term. For example

$$\begin{cases} \dot{x}_1 = 6x_1 + 10x_2 + 4x_3 + 20\epsilon x_1 - \xi_1\epsilon_1 \\ \dot{x}_2 = 10x_1 - 6x_3 + 20\epsilon x_2 - \xi_2\epsilon_1 \\ \dot{x}_3 = 4x_1 - 6x_2 + 2x_3 + 20\epsilon x_3 - \xi_3\epsilon_1 , \end{cases}$$

where

$$\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} = \mathbf{e}^1 \quad \text{and} \quad \epsilon_1 = \operatorname{sgn}(x_1\xi_1 + x_2\xi_2 + x_3\xi_3).$$

5. Results. The eigenvalues and eigenvectors of two matrices were computed at Project Cyclone, using REACs and Reeves multiplying servos. One of the matrices was the 3 × 3 listed above. The other was a 6 × 6 matrix, listed in reference 1 and reprinted below:

$$\begin{bmatrix} .06667 & .02634 & -.04640 & -.07368 & -.02131 & -.00431 \\ .02634 & .26841 & -.02243 & .15952 & -.05923 & -.12797 \\ -.04640 & -.02243 & .10932 & .05150 & -.04100 & .08558 \\ -.07368 & .15952 & .05150 & .25152 & -.01141 & -.07169 \\ -.02131 & -.05923 & -.04100 & -.01141 & .14403 & .01105 \\ -.00431 & -.12797 & .08558 & -.07169 & .01105 & .19450 \end{bmatrix}$$

A comparison between the computed and actual values of the eigenvector components are listed in Table 1.

The eigenvalues were not computed on the REAC, but a comparison may be made between $(A\mathbf{e}^i\cdot\mathbf{e}^i)/(\mathbf{e}^i\cdot\mathbf{e}^i)$ (where $\mathbf{e}^i$ is the computed eigenvector) and λ_i. This is shown in Table 2.

6. Additional Remarks. In the case of multiple eigenvalues, the eigenvectors are not, of course, unique. The method of this paper will then yield some set of orthonormal eigenvectors belonging to the eigenvalue. If two eigenvalues differ only slightly, the corresponding eigenvectors will be computed with less accuracy. But the corresponding eigenvalues will be obtained with no sacrifice in accuracy. In fact, the eigenvalues are relatively insensitive to errors in the eigenvectors. This is vividly demonstrated in Table 2, where the eigenvalues are seen to be correct to five or six significant figures.

The method of this paper can be extended to complex Hermitian matrices. The theorems listed in the mathematical formulation are basically still true if

FIG. 1.—Program for typical problem.

TABLE 1

Matrix 1

e^1		e^2		e^3	
computed	actual	computed	actual	computed	actual
.7933	.7930···	−.3189	−.3184···	−.5193	−.5194···
.6077	.6078···	.3562	.3555···	.7096	.7101···
−.0419	−.0415···	−.8794	−.8788···	.4762	.4754···

Matrix 2

e^1		e^2		e^3	
computed	actual	computed	actual	computed	actual
.0408	.0414··	−.3558	−.3554··	.3253	.3261··
−.6762	−.6763··	−.1227	−.1221··	.2627	.2639··
.0387	.0394··	.6012	.6014··	.2129	.2134··
−.5746	−.5745··	.5053	.5051··	−.3048	−.3044··
.1376	.1382··	−.0894	−.0889··	−.8014	−.8006··
.4361	.4361··	.4836	.4838··	.2104	.2116··

e^4		e^5		e^6	
computed	actual	computed	actual	computed	actual
−.3742	−.3732··	.2977	.2989··	.7326	.7328··
−.5208	−.5213··	−.3411	−.3406··	−.2653	−.2650··
.0686	.0678··	−.6013	−.6014··	.4741	.4743··
.0214	.0222··	.5271	.5271··	.2095	.2093··
−.4721	−.4728··	−.2775	−.2772··	.1771	.1777··
−.6004	−.6002··	.2789	.2799··	−.3039	−.3040··

TABLE 2

Matrix 1			Matrix 2		
	From Computed Eigenvector	Actual		From Computed Eigenvector	Actual
λ_1	6.727878	6.727881···	λ_1	.4982334	.4982344···
λ_2	2.938089	2.938091···	λ_2	.25946618	.25946632···
λ_3	−5.665964	−5.665972···	λ_3	.1759010	.1759013···
			λ_4	.0823481	.0823483···
			λ_5	.01581329	.01581310···
			λ_6	.00268708	.00268704···

one extends the definition of inner product to complex vectors by the formula $\mathbf{x}\cdot\mathbf{y} = \sum_{i=1}^{n} x_i\bar{y}_i$ [3].

The quadratic form $A\mathbf{x}\cdot\mathbf{x} = \sum_{i,j=1}^{n} a_{ij}x_i\bar{x}_j$ is real, and the extremum property can be used in the same way to find the real and imaginary parts of the eigenvectors.

Even normal matrices can be handled if one separates them into their real and imaginary parts [3].

7. Conclusion. The eigenvalue problem for real symmetric, or Hermitian, matrices can be solved on an electronic analog computer by formulating it as an extremum problem. Both the eigenvectors and eigenvalues can be obtained. With care, three place accuracy can be obtained for the eigenvector. If the eigenvalue is computed by hand from the eigenvector, six place accuracy can be obtained for the eigenvalues.

Reeves Instrument Corporation,
Garden City, New York

1. Olga Taussky, Editor, "Contributions to the solution of linear equations and the determination of eigenvalues," NBS *Applied Math. Series* 39, 1954, p. 60.

2. Landis Gephart, "Linear algebraic systems and the REAC," MTAC, v. 6, 1952, p. 190–203.

3. Paul R. Halmos, *Finite Dimensional Vector Spaces*, Princeton University Press, New Jersey, 1948, p. 124–134.

4. E. Bodewig, *Matrix Calculus*, North Holland Publishing Company, Amsterdam, 1956, p. 54–56 and 245.

5. Insley Pyne, "Linear programming on an electronic analogue computer," A.I.E.E. *Transactions Annual*, Part 1, Communication and Electronics, 1956, p. 139–143.

6. Clarence L. Johnson, *Analog Computer Techniques*, McGraw-Hill, New York, 1956.

7. Abraham Many, "Improved electrical network for determining eigenvalues and eigenvectors of a real symmetric matrix," *Review of Scientific Instruments*, v. 21, 1950, p. 972–974.

8. V. Rogla, "Analog machine for algebraic computations," presented at the International Analog Computation Meeting, Brussels, Belgium, 1955.

Graphical Evaluation of a Convolution Integral

By T. Mirsepassi

1. Introduction. The analytical expression which defines the response of a linear system to an arbitrary excitation is, in general, in the form of a convolution integral [1, 2, 3], i.e.,

$$(1)\quad x(t) = \int_0^t f(\tau)\cdot h(t-\tau)\,d\tau = \int_0^t f(t-\tau)\cdot h(\tau)\,d\tau = \int_0^t f(t-\tau)\cdot g'(\tau)\,d\tau$$

where:

$g(t)$ = response of system to unit step, and $g(0) = 0$,
$h(t) = g'(t)$ = response of system to unit impulse,
$f(t)$ = arbitrary excitation function,
$x(t)$ = response function of system to $f(t)$,
t = independent variable: time in the case of a time impulse such as in thermal [2] or electrical [3, 5] transients, and position in the case of space impulses, such as in deflection of beam or stretched cord under a space-variable load [4].

Usually the system—and therefore $g(t)$ or $h(t)$—is known and it is desired to find the response to a given excitation. Sometimes the excitation function is not known and is to be found from a given response. An example of this type occurs in the conduction of heat when, in a given solid, the transient temperature at a point inside the solid is recorded and the temperature-time function on the boundary (or in the ambient) is of interest. Whenever analytical integration of (1) is possible, numerical evaluation can be carried out easily; otherwise it becomes unusually lengthy and uneconomical. For such cases graphical treatment is advisable. The classical method [3] of graphical integration of (1) consists of:

1. Plotting $h(\tau)$, curve H in Fig. 1A

2. Folding $f(\tau)$, i.e., plotting it with $+\tau$ axis to the left, curve F in Fig. 1B

3. Translating Fig. 1B on Fig. 1A such that the τ axes coincide; for evaluating $x(t)$ at $t = t_1$, slide Fig. 1B on Fig. 1A until the origin of the τ axis in Fig. 1B falls on $\tau = t_1$ of the τ axis in Fig. 1A, see Fig. 1C.

4. Plotting the product curve, HF in Fig. 1C. Since HF represents $f(t_1 - \tau)\cdot h(\tau)$, the value of $x(t)$ for $t = t_1$ is shown by the area under this curve and between the two lines $\tau = 0$ and $\tau = t_1$. The numerical value of this area, when carried along the ordinate of $t = t_1$, locates one point of the response function, namely $x(t_1)$, P in Fig. 1D. Thus the mathematical process of convolution may be interpreted graphically by four operations: folding Fig. 1B, translating Fig. 1C, then multiplying and integrating. The above method is tedious—for each additional evaluation of $x(t)$ three operations (namely, translation, multiplication, and integration) must be repeated.

In this paper a new method is described which is based on a finite-difference form of (1) and graphical multiplication. By means of this method, evaluation of

Received December 4, 1957; in revised form, September 24, 1958.

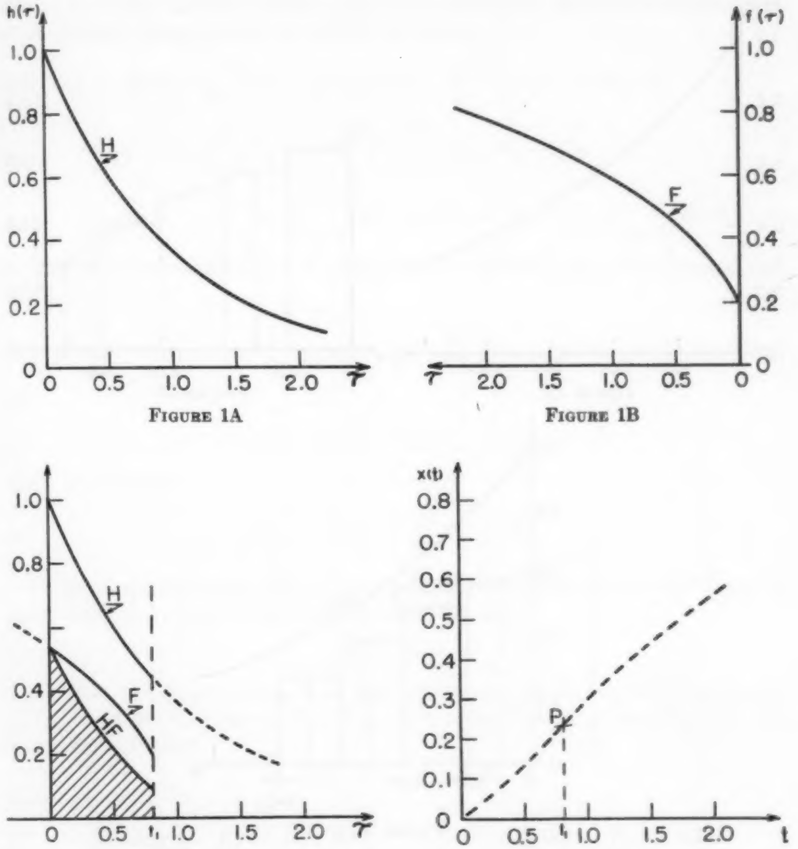

FIGURE 1A

FIGURE 1B

FIGURE 1C

FIGURE 1D

$x(t)$ at any value of t is reduced to adding a series of readings. The method becomes especially time-saving when, on a given system, (1) is to be solved a number of times.

2. Method. The function $h(\tau)$ is plotted in Fig. 2A. In Fig. 2B, drawn on transparent paper, folding of the excitation $f(\tau)$, drawn as a dotted line, is approximated by a step function drawn as a solid line; the duration of all steps along τ axis is the same and is denoted by $\Delta\tau$, and the magnitude of the steps is denoted by F_i, $(i = 1, 2, \cdots)$. The scales along the abscissa and ordinate in Fig. 2B have been taken, respectively, as equal to the scales along the abscissa and ordinate of Fig. 2A, and unit scale along the ordinate equals unit length. Now, superimpose Fig. 2B on Fig. 2A so that the horizontal axes coincide, see Fig. 2C, and that:

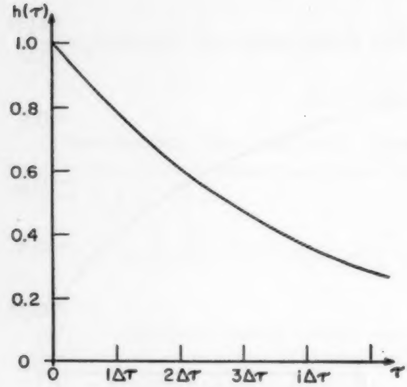

FIGURE 2A

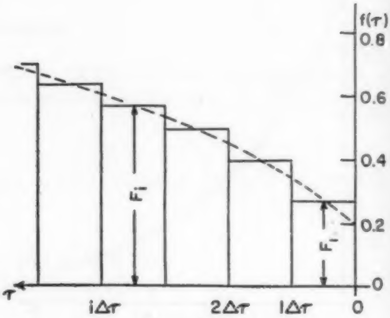

FIGURE 2B

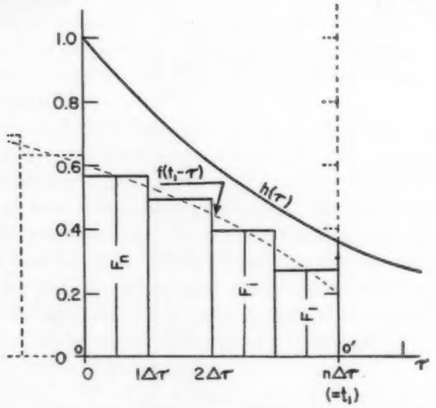

FIGURE 2C

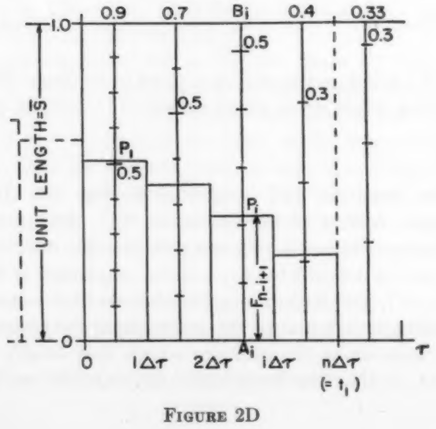

FIGURE 2D

$OO' = t_1 = n\Delta\tau$, t_1 being a known value of t at which the convolution integral is to be evaluated. Thus (1) may be written as follows:

$$\begin{aligned} x(t_1 = n\Delta\tau) &= \int_0^{t_1} f(t_1 - \tau)\cdot h(\tau)\,d\tau = \int_0^{n\Delta\tau} f(t_1 - \tau)\cdot h(\tau)\,d\tau \\ &= \int_0^{\Delta\tau} F_n\cdot h(\tau)\,d\tau + \int_{\Delta\tau}^{2\Delta\tau} F_{n-1}\cdot h(\tau)\,d\tau + \cdots \\ &\quad + \int_{(i-1)\Delta\tau}^{i\Delta\tau} F_{n-i+1}\cdot h(\tau)\,d\tau + \cdots + \int_{(n-1)\Delta\tau}^{n\Delta\tau} F_1\cdot h(\tau)\,d\tau. \end{aligned} \tag{2}$$

Since F_i remains constant in each integral, it may be taken out of the integral sign; therefore

$$x(n\Delta\tau) = \sum_{i=1}^{i=n} F_{n-i+1} \int_{(i-1)\Delta\tau}^{i\Delta\tau} h(\tau)\,d\tau = \sum_{i=1}^{i=n} F_{n-i+1}\,\{g(i\Delta\tau) - g[(i-1)\Delta\tau]\}. \tag{3}$$

Let

$$G_i = g(i\Delta\tau) - g[(i-1)\Delta\tau]. \tag{4}$$

Then (3) becomes

$$x(t_1 = n\Delta\tau) = \sum_{i=1}^{i=n} F_{n-i+1}\cdot G_i\,. \tag{5}$$

For graphical multiplication of $F_{n-i+1}\cdot G_i$, Fig. 2D is arranged. In this figure the center lines of $\Delta\tau$ intervals are graduated by the scales

$$\bar{S}_i = \bar{S}/G_i \tag{6}$$

where $\bar{S}$ = unit length. From these graduations one can read directly the product $F_{n-i+1}\cdot G_i$ at the intersection P_i, (Fig. 2D), of the step F_{n-i+1} and the center line A_iB_i. In fact, since

$$\overline{A_iP_i} = F_{n-i+1} \times \bar{S}$$

therefore:

$$\text{Reading at } \quad P_i = \overline{A_iP_i}/\bar{S}_i = (F_{n-i+1}\cdot\bar{S})/(\bar{S}/G_i) = F_{n-i+1}\cdot G_i\,. \tag{7}$$

Consequently (5) becomes

$$x(t_1 = n\Delta\tau) = \sum_{i=1}^{i=n} F_{n-i+1}\cdot G_i = \sum_{i=1}^{i=n} \quad \text{Reading at } P_i\,. \tag{8}$$

Thus, by means of Fig. 2D, the evaluation of $x(t)$ at $t_1 = n\Delta\tau$ is reduced to the addition of n readings.

The evaluation of $x(t)$ for other values of t is done similarly by sliding the transparent Fig. 2B on Fig. 2D to the new position and adding the readings at the new intersections of the steps and the center lines.

3. Notes.

1. The stepwise approximation of the excitation function may be omitted whenever the function is approximately linear inside each interval. In fact, the average step, if drawn, will intersect the center line at about the same point as the function itself. The error due to stepwise approximation, however, will exist.

2. The response at $t = n\Delta\tau$ to unit step applied at $t = 0$ is

$$g(n\Delta\tau) = \sum_{i=1}^{i=n} G_i = \sum_{i=1}^{i=n} \text{Reading at } B_i \tag{9}$$

where B_i is the point of intersection at which the unit ordinate line intersects the center line of the $i^{\text{th}}\Delta\tau$ interval, Fig. 2D.

3. The response at $t = (i - \frac{1}{2})\Delta\tau$ to unit impulse applied at $t = 0$, is

$$G_i/\Delta\tau = (\text{Reading at } B_i)/\Delta\tau. \tag{10}$$

In fact, from (4)

$$G_i/\Delta\tau = \{g(i\Delta\tau) - g[(i-1)\Delta\tau]\}/\Delta\tau.$$

The right side of this equation is the central difference of $g'(t)$ and thus represents:

$$g'(t) = h(t)$$

at

$$t = (i - \tfrac{1}{2})\Delta\tau.$$

Fig. 2D, which represents values proportional to the impulsive response, henceforth will be referred to as the "unit impulse chart".

4. In problems concerning the conduction of heat, it is sometimes of interest to know the rate of temperature change at a point inside a solid when the surface (or ambient) temperature varies with time (e.g., quenching). This graphical technique can also be employed for problems of this type, i.e., problems where the derivatives of $x(t)$ are of interest. In fact, using the Leibnitz rule [6]:

$$x'(t) = \frac{d}{dt}\left[\int_0^t f(\tau)\cdot h(t-\tau)\,d\tau\right] = \int_0^t f(\tau)\cdot h'(t-\tau)\,d\tau + f(t)\cdot h(0)$$

since $g(t)$ represents the response at a known point inside the solid, which is initially at a steady state, when its surface (or ambient) temperature has undergone a unit step change—i.e., since: $g'(t) = h\ (t) = 0$ for $t = 0$; therefore

$$x'(t) = \int_0^t f(\tau)h'(t-\tau)\,d\tau = \int_0^t f(t-\tau)\cdot h'(\tau)\,d\tau. \tag{11}$$

In this case, (4) becomes

$$H_i = h(i\Delta\tau) - h[(i-1)\Delta\tau]; \tag{12}$$

and, as explained before, a chart can be arranged similar to Fig. 2D. By a discussion analogous to that of Note 2, the resulting chart may be referred to as the "unit doublet chart".

5. In complicated systems, $g(t)$, $h(t)$, or $h'(t)$ may be obtained experimentally or, possibly, by an analog computer. Although, in such systems, response to an arbitrary excitation may be found similarly, it often pays to make a unit-step run, arrange the "unit impulse chart", and find $x(t)$ graphically. This has two advantages:

(a) Considerable saving of computer time

(b) Possibility of solving additional response problems when the experimental setup or analog computer is no longer available.

4. Example. A linear system is defined by its response to the unit step according to the following relationship

$$g(t) = 1 - e^{-t}. \tag{13}$$

A. RESPONSE TO A GIVEN EXCITATION

Find the response of this linear system to

$$f(t) = t, \qquad 0 \leqq t \leqq 1$$

$$f(t) = 1, \qquad 1 \leqq t \leqq 2.$$

1. *Preparation of "unit impulse chart"*

Equation (4) becomes:

$$G_i = [1 - e^{-i\Delta\tau}] - [1 - e^{-(i-1)\Delta\tau}] = e^{-(i-1)\Delta\tau} - e^{-i\Delta\tau}. \tag{14}$$

With due consideration for the required accuracy, select $\Delta\tau$ for example, take $\Delta\tau = 0.2$ and calculate Table 1.

TABLE 1
Calculation of $1/G_i$ from Equation (14).

i	1	2	3	4	5	6	7	8	9	10
$i\Delta\tau$	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
$e^{-(i-1)\Delta\tau}$	1.000	0.8187	0.670	0.549	0.4493	0.3679	0.3012	0.2466	0.2019	0.1653
$e^{-i\Delta\tau}$	0.819	0.6703	0.549	0.449	0.3679	0.3012	0.2466	0.2019	0.1653	0.1353
G_i	0.181	0.148	0.121	0.100	0.0814	0.0667	0.0546	0.04470	0.0366	0.0300
$1/G_i$	5.525	6.757	8.264	10.000	12.285	14.992	18.315	22.371	27.322	33.333

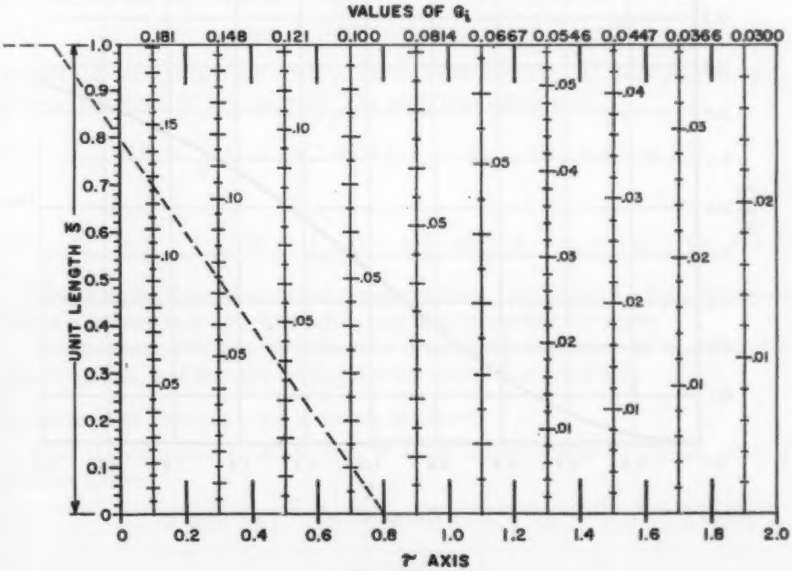

FIGURE 3A

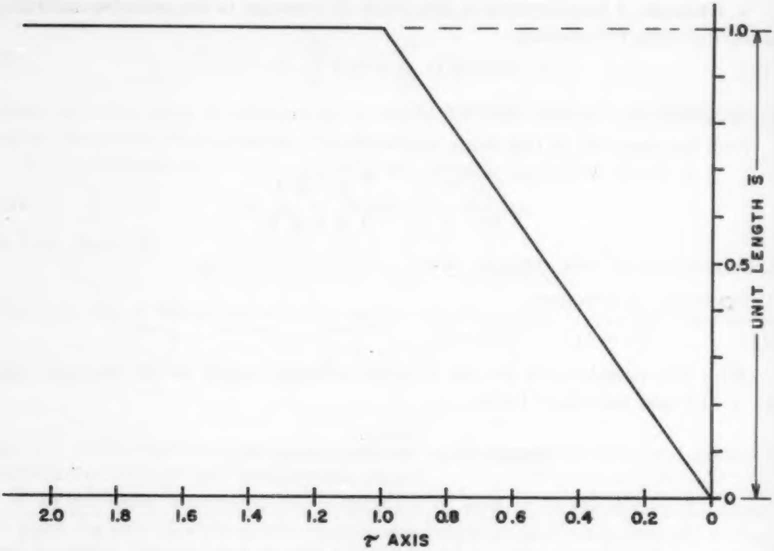

Figure 3B

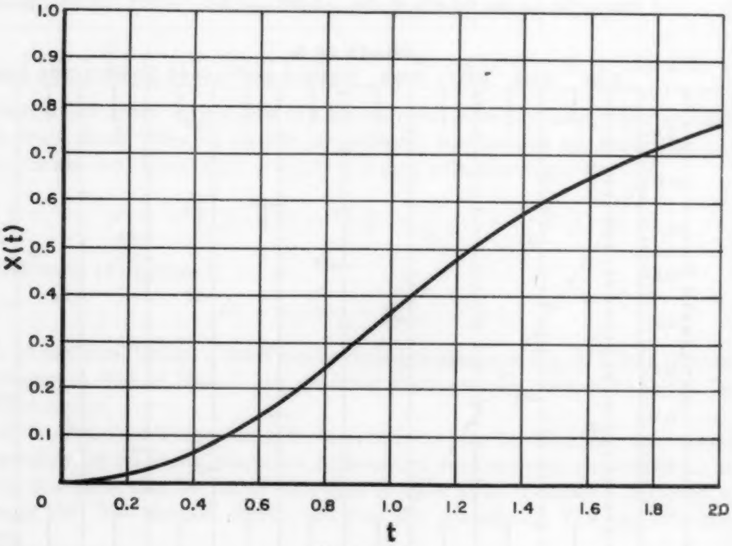

Figure 3C

TABLE 2

Comparison of values of response obtained by evaluation of Equations (15) and (16), finite-difference Equation (8), and actual readings of chart.

t Response	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
Theoretical: Equations (15) & (16)	0.0187	0.0703	0.1488	0.2493	0.3679	0.4824	0.5763	0.6531	0.7159	0.7675
Finite-Diff. Equation (8)	0.0181	0.0692	0.1473	0.2475	0.3658	0.4808	0.5750	0.6522	0.7155	0.7670
Error (%)	−3.2	−1.5	−1.0	−0.7	−0.6	−0.3	−0.2	−0.1	−0.05	−0.01
Reading of Chart	0.018	0.069	0.147	0.248	0.367	0.480	0.575	0.651	0.714	0.766
Error (%)	−3.7	−1.8	−1.2	−0.5	−0.2	−0.5	−0.2	−0.3	−0.2	−0.2
Graphical Error (%)*	−0.5	−0.3	−0.2	+0.2	+0.4	−0.2	0	−0.2	−0.15	−0.19

* "Graphical error" = "Reading of Chart" error − "Finite Difference" error.

Mark $\Delta\tau$ intervals on the τ axis in Fig. 3A; use an arbitrary scale.

Select a unit length, $\bar{S}$, (Fig. 3A), and graduate each center line of $\Delta\tau$ intervals by the corresponding unit scale $\bar{S}_i$, where

$$\bar{S}_i = (1/G_i)\bar{S}.$$

For example, for $i = 1$ (the first center line), $\bar{S}_1 = (1/0.181)\bar{S} = 5.525\bar{S}$, or $0.1\bar{S}_1 = 0.5525 \cdot \bar{S}$.

2. *Evaluation of response*

Plot the folding of $f(t)$, take the same $\Delta\tau$ intervals as in Fig. 3A, and use unit scale $\bar{S}$ for the ordinate; see Fig. 3B which should be drawn on transparent paper.

Superimpose Fig. 3B on Fig. 3A, as is shown by the dotted line in Fig. 3A for $t = 0.8$, and add the readings; thus, for $t = 0.8$, the response is:

$$x(t) = 0.010 + 0.036 + 0.076 + 0.127 = 0.248.$$

Repeat this second step for other values of t and plot Fig. 3C. It is interesting to analyze the errors in this example. The analytical solution is:

$$x(t) = \int_0^t (t - \tau)e^{-\tau}\, d\tau = t + e^{-t} + 1, \quad \text{for} \quad 0 \leqq t \leqq 1 \tag{15}$$

and

$$x(t) = \int_0^{t-1} 1 \cdot e^{-\tau}\, d\tau + \int_{t-1}^{t} (t - \tau)e^{-\tau}\, d\tau = 1 - (e - 1)e^{-\tau}, \quad \text{for} \quad t \geqq 1. \tag{16}$$

Table 2 shows theoretical values computed from (15) and (16), finite difference values computed from (8), and values obtained by reading the chart.

It is interesting to note that the error is primarily introduced by the difference approximation, and that the graphical error does not exceed 0.5%.

B. EXCITATION FUNCTION FOR A GIVEN RESPONSE

Find the excitation function, $f(t)$, if in the above system a response, $x(t)$, is given as follows:

t	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
$x(t)$	0	0.018	0.069	0.147	0.248	0.367	0.480	0.575	0.651	0.714	0.766

Evaluation of excitation function

Plot $x(\tau)$ with the τ axis increasing from right to left with the abscissa scale the same as in Fig. 3A, (i.e., the distance between two center lines representing $\Delta\tau = 0.2$), and the ordinate scale equal to the unit length $\bar{S}$ (Fig. 4A, which should be drawn on transparent paper).

Superimpose Fig. 4A on Fig. 3A as shown in enlarged view in Fig. 4B. Let *A*

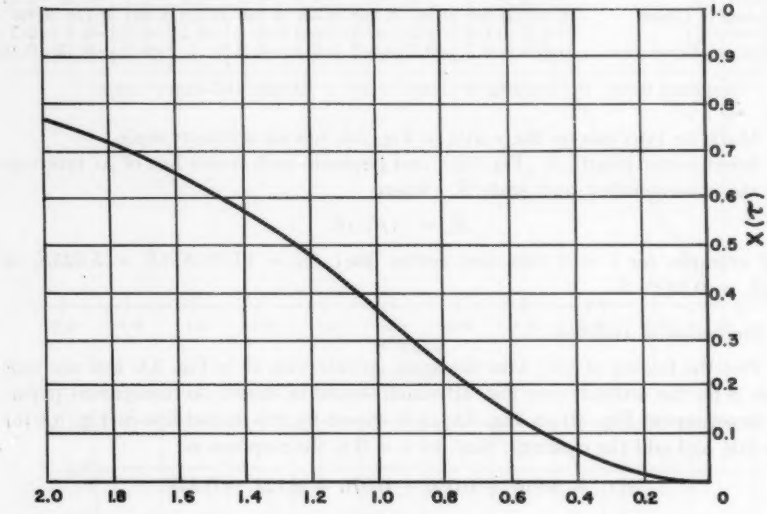

FIGURE 4A

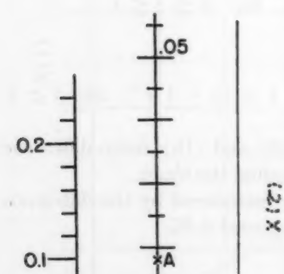

FIGURE 4B

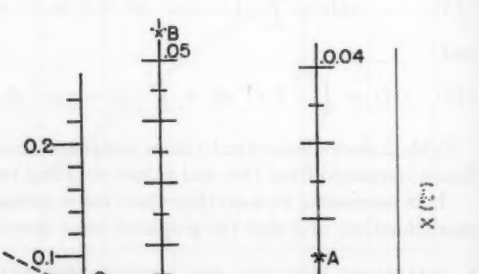

FIGURE 4C

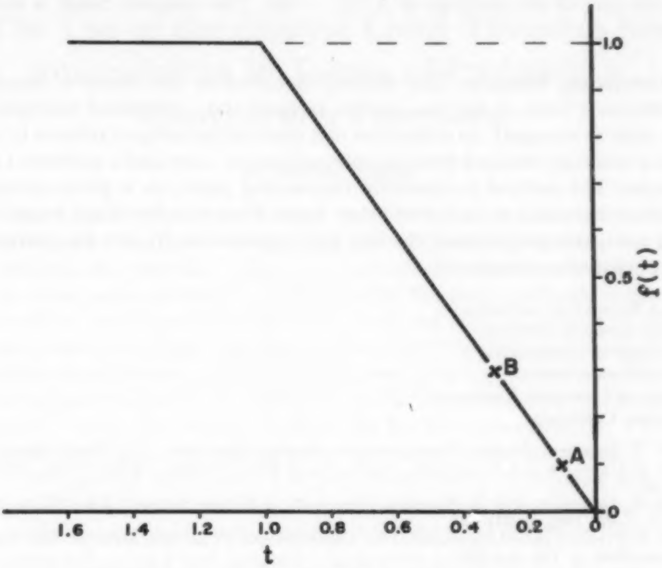

FIGURE 4D

denote the intersection of the unknown $f(t)$ with the center line. From (5) written for $n = 1$,

$$x(\Delta\tau) = F_1 \cdot G_1 .$$

The lefthand side of this equation is the reading at P from the graduated unit scale. The righthand side is the reading at the point A from the graduated center line. The lefthand side is known—e.g., 0.018 in Fig. 4B. The righthand side must then equal the reading at P. Therefore, the location of A is established in this case as the point 0.018 (read from the graduations of the center line) along the center line.

Then, slide Fig. 4A on Fig. 3A to the second position, as shown in Fig. 4C. Let B denote the new intersection of $f(t)$ and the first center line. From (5), written for $n = 2$:

$$x(t) = F_2 \cdot G_1 + F_1 \cdot G_2 .$$

Since: $x(t)$ = Reading at Q from graduated unit scale (i.e., 0.069 in Fig. 4C) and $F_1 \cdot G_2$ = Reading at A from graduated center line (i.e., 0.0148 in Fig. 4C), therefore:

$$F_2 \cdot G_1 = \text{Reading at } B \text{ from the graduated center line}$$

$$= x(t) - F_1 \cdot G_2 = 0.0690 - 0.0148 = 0.0542$$

and in Fig. 4C, Point B is located where the reading along the center line is 0.0542. Thus, after each translation, the reading at the new intersection of $F(t)$ and the first center line is obtained by subtracting from $x(t)$ (read from graduated unit scale) the sum of the readings at A, B, $\cdots$ etc. The complete result is shown in Fig. 4D.

5. Concluding Remarks. The method described in this paper is based on a finite-difference form of the convolution integral and a graphical multiplication. Once a chart is arranged, an evaluation of a convolution integral reduces to adding a series of readings obtained from the intersections of a line and a number of graduated scales. The method is especially time-saving when, on a given system, the convolution integral is to be solved many times. Heat-transfer charts based on this method are under preparation; the first part appeared in [7] and the continuation will be published as completed.

Heat and Mass Flow Laboratory,
Corporate Research Division,
Aerojet-General Corporation,
Azusa, California; and
University of California Extension,
Los Angeles, California

1. W. T. THOMSON, *Laplace Transformation*, Prentice Hall, Inc., New York, 1950, p. 37–38.
2. H. S. CARSLAW & J. C. JAEGER, *Conduction of Heat in Solids*, Oxford Press, New York, 1950, p. 20.
3. M. F. GARDNER & J. L. BARNES, *Transients in Linear Systems*, John Wiley & Sons, Inc., New York, 1942, p. 231–234.
4. L. A. PIPES, *Applied Mathematics for Engineers and Physicists*, McGraw-Hill Book Co., New York, 1946, p. 213 and 222.
5. STANFORD GOLDMAN, *Transformation Calculus and Electrical Transients*, Prentice Hall, Inc., New York, 1950.
6. C. R. WYLIE, JR., *Advanced Engineering Mathematics*, McGraw-Hill Book Co., Inc., New York, 1951 p. 591.
7. T. J. MIRSEPASSI, "Heat-transfer charts for time-variable boundary conditions—semi-infinite solid," *British Chemical Engineering*, March 1959, p. 130–136.

The Use of the Central Limit Theorem for Interpolating in Tables of Probability Distribution Functions

By Gerard Salton

In using tables of probability density and distribution functions, the difficulty of interpolating for functional values which are not directly tabulated constitutes a major problem. In particular, when the variance of a random variable becomes small, the corresponding density and distribution functions approach, respectively, a delta-function and a step-function. As a result, for small variations in the variables, there occur very large variations in the corresponding functional values, and interpolation by difference methods will give very poor approximations. Moreover, space limitations in a volume of tables often make it impractical to tabulate a given function on a mesh fine enough to allow for accurate interpolation. This often increases the difficulty of interpolating for a given probability distribution.

By a fundamental limit theorem of probability theory it is known that, under rather general conditions, the sum of a set of n independent random variables appropriately standardized by the mean and the standard deviation is asymptotically normal with mean 0 and variance 1, as n tends to infinity [1]. A method which makes use of this property to improve the accuracy of interpolating in tables of probability distribution functions was suggested by Rossow [2]. The binomial probability distribution is chosen here for purposes of illustration. The method described is however applicable to any set of random variables which obeys the central limit theorem.

The cumulative binomial probability distribution may be denoted by

$$E(n, r, p) = \sum_{i=r}^{n} C_i^n (1 - p)^{n-i} p^i$$

where $0 \leqq p \leqq 1$ and $0 \leqq r \leqq n$. If one considers a series of n independent repetitions of some random experiment, then $E(n, r, p)$ represents the probability of at least r successes in n repetitions of the experiment, p being the probability of success for each experiment. By the theorem previously stated, it is known that for a given value of p, and for increasing values of n

$$E(n, r, p) \approx \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}\, dt = 1 - N(x)$$

where $x = (r - np - \frac{1}{2})/\sqrt{np(1 - p)}$, and $N(x)$ denotes the cumulative standard normal distribution. The effectiveness of the approximation increases with increasing values of $\sqrt{np(1 - p)}$, so that for a given value of n, the approximation is closest for $p = \frac{1}{2}$. The argument x, corresponding to a given value of $N(x)$, is often called the normal deviate corresponding to the total frequency N.

Consider now the problem of interpolating in a table of the binomial probability

Received April 9, 1958.

distribution. As an example, let it be desired to find some value $E(n_0, r_0, p_0 + \Delta p)$, where the function is not tabulated for the argument $p_0 + \Delta p$. The standard procedure consists in taking the values of $E(n, r, p)$ corresponding to the arguments $n_0 r_0 p_0$, $n_0 r_0 p_1$, $\cdots$, $n_0 r_0 p_{k+1}$ and in constructing a $k + 1^{\text{th}}$ degree polynomial which takes on the given functional values at the $k + 2$ given points. The value of the polynomial at the point $n_0 r_0 p_0 + \Delta p$ is then calculated and taken to be equal to $E(n_0, r_0, p_0 + \Delta p)$.

It is proposed to modify this procedure by making use of the fact that the binomial distribution is often much better approximated by the corresponding normal distribution than by the interpolation polynomial which is constructed in the standard method. To this effect, it is first assumed that $E(n_0, r_0, p_0) = 1 - N(x_0)$, $E(n_0, r_0, p_1) = 1 - N(x_1)$, $\cdots$, $E(n_0, r_0, p_{k+1}) = 1 - N(x_{k+1})$. The $k + 1$ normal deviates corresponding to the $k + 1$ values of $1 - N$ are then found in a table of the normal distribution. Thereafter the interpolation is performed in the normal deviates, that is, from the values $-x_i$ corresponding to the arguments p_i, a value $-x_p$ is determined which corresponds to the argument $p_0 + \Delta p$. The value of $1 - N(x_p)$ is then looked up in a table of the normal distribution, and $E(n_0, r_0, p_0 + \Delta p)$ is approximated by $1 - N(x_p)$.[*] The procedure is outlined in the following diagram, where the arrows denote a table look-up and the brace stands for the evaluation of an ordinary interpolation polynomial:

$$\left.\begin{array}{lcccccccc}
p_0 & \rightarrow & E(n_0, r_0, p_0) & = & 1 - N(x_0) & \rightarrow & -x_0 \\
p_1 & \rightarrow & E(n_0, r_0, p_1) & = & 1 - N(x_1) & \rightarrow & -x_1 \\
\vdots & & \vdots & & \vdots & & \vdots \\
p_{k+1} & \rightarrow & E(n_0, r_0, p_{k+1}) & = & 1 - N(x_{k+1}) & \rightarrow & -x_{k+1}
\end{array}\right\} -x(p_0 + \Delta p) \rightarrow 1 - N(x_p)$$

$$= E(n_0, r_0, p_0 + \Delta p).$$

The procedure is easier to perform if tables are available which tabulate the normal deviates as a function of the cumulative normal probabilities [3]. However, tables which give $N(x)$ as a function of x can also be used [4]. In either case, the accuracy of the method depends on the availability of accurate values of the normal deviates; this is especially important for cumulative probabilities close to 0 and close to 1 where the deviates change rapidly for small changes in the probabilities. It should be noted that the additional labor required by the proposed method is restricted to some table look-up operations which can be performed rapidly.

A study was made of the efficiency of the method, based on a recent tabulation of the cumulative binomial probability distribution [5]. It is stated in that volume that interpolation by standard difference methods gives poor results for certain ranges of the arguments. In particular, p-wise interpolation is poor for $n > 100$ and p small, r-wise interpolation is poor for $n < 100$ and p small, and n-wise interpolation is inaccurate for $n > 100$ and p close to $\frac{1}{2}$.

The normal deviate method fills the need for both p-wise and n-wise interpola-

[*] Since the arguments of $1 - N(x)$ are the negatives of the corresponding arguments of $N(x)$, the method can also be carried out by interpolating in the values x_i, rather than in the values $-x_i$. Such an interpolation will result in the determination of the function $N(x_p)$ corresponding to $1 - E(n_0, r_0, p_0 + \Delta p)$.

tion, since it is most accurate for large n where the standard method fails. It can also be used for interpolation in r when n is not too small. When interpolating for p and for n, a four-point interpolation in the normal deviates was found to result in final probabilities whose error did not exceed 5.10^{-5}, except for very small p ($p \leq 0.03$) where the error could be as high as 5.10^{-3}. Four-point interpolation in r for $n > 30$ resulted in errors not exceeding 5.10^{-4}.

A different method which also improves the interpolation in tables of probability distribution functions consists in approximating the given distribution by the logistic function [6]

$$L(x) = \frac{1}{1 + e^{-\beta x}}$$

in lieu of the cumulative standard normal distribution. In fact it is known that

$$L(x) \approx N(x),$$

the two functions differing by less than 0.01 when β is close to 1.7. The interpolation proceeds exactly as before except that $L(x)$ is used for $N(x)$. The arguments x corresponding to various values of $L(x)$ are sometimes called logits. By inverting the expression for $L(x)$ it is seen that

$$x = \frac{1}{\beta} \ln \frac{L}{1 - L}.$$

Tables of the natural logarithms and exponential tables can therefore be used, if logit tables are not available [6].

Linear interpolation in the logits will sometimes give slightly better values than the corresponding interpolation in the normal deviates. However, a higher order interpolation formula brings much less improvement in the logits than it does in the normal deviates. A four-point interpolation in the normal deviates is usually to be preferred to the corresponding four-point formula in logits. It is again important to use accurate logit tables, especially for values of L close to 0 and 1. Since tables of the logistic function may not be generally available, interpolation in the normal deviates is usually easier to perform and will moreover give more accurate results for higher order interpolation formulas.

The following examples give a good idea of the improvement which can be

Comparison of Interpolation Methods

	E (600,225,0.375) p-wise, $\Delta p = 0.5$	E (1000,53,0.0625) p-wise, $\Delta p = 0.25$	E (39,6,0.08) r-wise, $\Delta r = 1$	E (689,285,0.42) n-wise, $\Delta n = 39$
Correct Values	.51541	.90685	.08786	.64612
Ordinary 2-pt. Ordinary 4-pt.	.51475 (−66) .51531 (−10)	.87776 (−2909) .90178 (−507)	.11597 (+2811) .09194 (+408)	.63732 (−880) .64157 (−455)
Normal Deviates 2-pt.	.51524 (−17)	.90497 (−188)	.09004 (+218)	.64263 (−349)
Normal Deviates 4-pt.	.51542 (+1)	.90685 (0)	.08770 (−16)	.64616 (+4)
Logits 2-pt. Logits 4-pt.	.51537 (−4) .51543 (+2)	.91249 (+564) .90973 (+288)	.08456 (−330) .08726 (−60)	.64401 (−211) .64870 (+258)

achieved over the standard interpolation method by using either normal deviates or logits. In each case, the error in the value obtained is shown in parentheses.

Computation Laboratory, Harvard University,
Cambridge, Massachusetts

1. Harald Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1951, p. 213–220.

2. E. Rossow, Technische Universität, Berlin, private communication.

3. R. A. Fisher & F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd Ltd., Edinburgh, 1948.

4. Harvard University, Computation Laboratory, *Annals*, v. 23: *Tables of the Error Function and of its First Twenty Derivatives*, Harvard University Press, Cambridge, Mass., 1952.

5. Harvard University, Computation Laboratory, *Annals*, v. 35: *Tables of the Cumulative Binomial Distribution*, Harvard University Press, Cambridge, Mass., 1955.

6. Joseph Berkson, "A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function," *Journal of the American Statistical Association*, v. 48, No. 263, September, 1953.

REVIEWS AND DESCRIPTIONS OF TABLES AND BOOKS

34[**A, B, C, D**].—L. Flavien, *Nouvelles Tables Numériques pour les Fonctions Usuelles de l'Analyse*, Gauthier-Villars, Paris, 1958, 63 p., 21 cm. Price $0.85.

This little volume has been compiled for students who are preparing "aux grands écoles scientifiques." It has been edited to conform with the 1956 program published by the *ministère de l'Education Nationale*. Tables included are:

I. Avertissement.
II. Values of n^2, n^3, and n^{-1}, n^{-2}, n^{-3}, $\sqrt{n}$ $\sqrt[3]{n}$ to 4D, for $n = 1(1)1000$.
III. Mantissa of $\log_{10} x$ to 4D for $x = 1(1)999$.
IV. $\text{Log}_e\, x$ to 4D for $x = 1(1)1000$.
V. 10^x to 4D for $x = 0(.01)1$.
VI. Sin x, tan x, ctn x, cos x to 4D, in degrees for $x = 0(0°.1)90°$.
VII. Sin x, tan x, ctn x, cos x to 4D, in grades for $x = 0(0^g.1)100^g$.
VIII. Arc sin x in radians to 5D for $x = 0(0.01)1$. Arc tan x in radians to 5D for $x = 0(0.01)1$ (var.)100.
IX. Radians r: to degrees (decimal), to grades (decimal), to 4D for $r = 1(1)10$, Radians to degrees, minutes, and seconds for $r = 0.1(0.1)0.9, 0.01(0.01)0.09. 0.001(0.001)0.009, 0.0001(0.0001)0.0009$.
X. Degrees d, minutes m, seconds s: to grades (4D), to radians (5D) for $d = 1(1°)90°$, $m = 1(1')60'$, $s = 1(1'')60''$.
XI. Grades G: to radians (4D), to (decimal) degrees (1D), to degrees, minutes, for $G = 1(1^g)100^g$. Centigrades C to minutes, seconds, (1D), for $C = 1(1^c)100^c$.
XII. "Remarkable" numbers (e.g., π, e, $n!$) and their common logarithms.

The Tables are well designed and easy to read.

Richard S. Burington

Bureau of Ordnance,
Navy Department,
Washington, District of Columbia

35[**A, B, I, N**].—P. Montagne, *Tables abrègées de puissances entières*, Dunod, Paris, 1958, xv + 411 p. + loose appendix 32 p., 28 cm. Price 5600 francs.

These extensive tables are designed to facilitate the evaluation of x^n, where n is a positive integer, more accurately than is possible with eight-place logarithmic tables. The main tables are divided into three sections comprising tables referred to as small, medium and large. The values of x^n are given in the small and medium tables to 15S, and in the large tables to 10–11S. Numerous special signs attached to the last figure give some information about the next place. There are no differences. The values of the arguments are:

Small tables (a), pages 3–17: $x = .2(.1)2$, $n = 0(1)m$, where m depends on x in the following way:

x	.2	.3	.4	.5(.1).9	1.1(.1)2
m	600	400	250	200	150

Small tables (b), pages 18-32: $x = .1(.01)1.53$, $n = 2(1)78$.

Medium tables, pages 64–157: $x = .097(.001)1.263$, $n = 2(1)26$.

Large tables, pages 160–411: $x = 0(.0001)1.2603$, $n = 2(1)10$.

In all tables except small (a), x is the vertical argument, and there is an overlap of about 7 lines between pages, instead of the usual one line or even none. This accounts for initial and terminal arguments ending in 7 and 3; thus, the medium tables are basically for $x = .1(.001)1.26$, the three extra arguments at each end being added for convenience in interpolation and so forth. The tables are photographically reproduced from manuscript written by professional calligraphers.

A loose appendix contains 29 pages of *Tables annexes* and 3 pages reserved for notes. The varied contents defy brief description in full. About half the pages contain matter relating primarily to the main tables, such as illustrative diagrams, indexes, and tables indicating the number of differences needed in interpolation. The remaining pages contain matter of more general applicability. Outstanding are two tables of Everett interpolation coefficients, one giving exact coefficients of second and fourth differences at interval 0.002, and the other, phenomenally extensive, giving at interval 0.01 coefficients to 12–13D of all even differences up to the sixteenth; except in the case of coefficients of the 14th and 16th differences, even differences (up to the second or fourth) of the interpolation coefficients are also given. There are also rather extensive tables of factorials, reciprocals of factorials, binomial coefficients, and useful constants (including Bernoulli and Euler numbers).

There are few references; it seems a pity not to mention the important British Association tables of powers [1]. The tables of x^n will certainly be found useful for tabular arguments, but when much interpolation would be needed, the use of logarithms as an alternative may still appeal on occasion. Obviously much work has gone into checking the tables, but a copious page of errata is pasted in, and it is stated that the last check revealed less than one error pĕr ten pages. There do exist, however, radix and similar special logarithmic tables which are slim, simple, and of absolute accuracy. Nevertheless, anyone much concerned with integral powers should consider what possibilities this volume may have for his purpose. An excellent feature of that part of the large tables which relates to the range $0 \leqq x \leqq 1$ is that, for given x, all the powers of x and $1 - x$ are contained on one line running right across two pages visible at an opening; this facilitates the solution of equations of the form $x^\alpha(1 - x)^\beta = k$, which occur in connection with chemical equilibria. The extensions for $x > 1$ may be found useful in connection with problems on compound interest.

Editions in English and several other languages are in preparation, as also is a table of fourth powers.

A.F.

[1] British Association for the Advancement of Science, *Mathematical Tables, vol. IX: Table of Powers giving Integral Powers of Integers*, Cambridge University Press, 1940. See MTAC, Review **169**, v. 1, 1945, p. 355–356.

36[G].—L. Lunelli & M. Sce, "Sulla ricerca dei k-archi completi mediante una calcolatrice elettronics," *Atti Convegno Intern. Reticoli e Geometrie Proiettive*, Palermo-Messina, 1957 (Roma 1958), p. 81–86.

[G].—L. Lunelli & M. Sce, *k-archi completi nei piani proiettivi desarguesiani di rango 8 e 16*, Politecnico di Milano, Centro di Calcolo Numerici, Milano 1958, 11 p.

The authors have programmed a CRC 102 A/P to search for complete k-arcs in a finite projective plane coordinatized by a Galois field. A k-arc is a set of k points such that no three lie on a line. It is complete if it is not contained in a $(k + 1)$-arc. To eliminate the most obvious duplications, only k-arcs through the four fundamental points $(0, 0, 1)$, $(0, 1, 0)$, $(1, 0, 0)$, $(1, 1, 1)$ were considered, but the authors regard as distinct k-arcs those which result from one another under collineations permuting the four points.

The procedures in the first paper are limited to planes of prime order. A description of the program is given and the results of a complete search of the plane of order seven are announced. Examples of complete k-arcs are given for planes of orders 11, 13, and 17.

In the case of the plane of order seven, 40 6-arcs are tabulated.

In the second paper, the program is extended to utilize the irreducible polynomials $x^3 + x + 1$ and $x^4 + x + 1$ to calculate in terms of the coordinatization by GF(8) and GF(16) for planes of these orders. (A misapprehension as to the general suitability of polynomials $x^n + x + 1$ does not affect the results in these cases.) Ten 10-arcs are listed for the plane of order eight, and 45 6-arcs. If permutations were considered, the authors could have reduced the tabulation to three 10-arcs and fifteen 6-arcs. Indeed, as they point out, the case of the 6-arcs can be reduced to the statement: any 5-arc can be completed to a complete 6-arc in precisely three ways. For the case 16 no exhaustive tabulation is given. Some 18 10-arcs, one 11-arc, two 12-arcs and three 18-arcs are listed. The last do not contain conics (ovals); this answers a question of B. Segre as to the existence of such arcs.

J. D. SWIFT

University of California,
Los Angeles, California

37[H, I, S].—R. L. CHAMBERS & E. V. SOMERS, *Solution of a radiation type non-linear differential equation*, Scientific Paper 8-0529-P6, Westinghouse Research Laboratories, Pittsburgh, Pennsylvania, October 1958.

The treatment of a heat transfer problem [1, 2] dealing with radiation from a cooling fin gives rise to a second order non-linear differential equation

$$\frac{d^2\theta}{dr^2} + \frac{1}{r + 1/(\rho - 1)}\frac{d\theta}{dr} - \gamma\theta^4 = 0;$$

with boundary conditions $\theta = 0$ at $r = 0$, $d\theta/dr = 0$ at $r = 1$. Tabulated results are presented from the numerical solution of the above equation over a range of values of r from 0 to 1.0, for the parametric values of $\rho = 1.001, 1.1, 1.25, 1.5, 2.0$, and 3.0 and $\gamma = 0$ to 4. The tables were computed on an IBM 704 digital computer by a finite-difference method. A copy has been deposited in the UMT file.

H.P.

[1] ROBERT L. CHAMBERS, *The determination of radiation fin efficiency and temperature distribution for one-dimensional heat flow in a circular fin*, University of Pittsburgh thesis, 1958.

[2] ROBERT L. CHAMBERS & E. V. SOMERS, *Radiation fin efficiency for one-dimensional heat flow in a circular fin*, Westinghouse Scientific Paper 8-0529-P4, 1958.

38[I, X, Z].—R. M. PEARCE, *Digital Computer Solution of the Two Group Diffusion Equations in Cartesian or Cylindrical Geometry with Application to the Datatron*, Report AECL No. 487, 1957, 27 p., 27 c.m. Available from Scientific Document Distribution Office, Atomic Energy of Canada Limited, Chalk River, Ontario, Canada. Price $1.00.

This report derives a set of difference equations for use in solving the two-group diffusion equations in XY or RZ geometries, and describes a Datatron program to solve the difference equations. The Datatron program and the output from a sample problem are included.

The Datatron program can handle problems with a maximum of 324 mesh points. This is less than some other currently available programs can handle.

The boundaries may be black (the value of the flux is forced to zero) lines of symmetry, or what is referred to in the report as reflector termination. In the case of reflector termination, an attempt is made to simulate the effects of a partially reflecting material bounding the cell.

The iteration scheme uses point relaxation. The latest values are used except in one case where their use would require extra machine time per iteration. No attempt is made to accelerate the convergence of the iteration scheme, even though there are several methods available which have been used successfully in similar programs.

Charles Dawson

Applied Mathematics Laboratory,
David Taylor Model Basin,
Washington 7, District of Columbia

39[L].—H. V. McIntosh, A. Kleppner & D. F. Minner, *Tables of the Herglotz Polynomials of Orders $\frac{3}{2}$–$\frac{8}{2}$ Transformation Coefficients for Spherical Harmonics*, Ballistic Research Laboratories, Memorandum Report No. 1097, Aberdeen Proving Ground, Maryland, 1957, 162 p., 28 cm.

The Herglotz polynomials H_{jk}^{n} (R) are the matrix elements of the unitary irreducible representations of the three-dimensional rotation group and are defined by:

$$H_{jk}^{n}(\sigma, \phi, \gamma) = (-1)^{j-k}e^{ij\gamma}e^{ik\alpha}(\cos \phi/2)^{2n-j+k}(\sin \phi/2)^{j-k}$$
$$\cdot \sum_{s} \frac{(-1)^{s} \tan^{2s} \phi/2}{(n+k-s)!(j-k+s)!(n-j-s)!\, s!},$$

where α, ϕ, γ are the Euler angles of the rotation R. These polynomials have the useful property that if R takes the coordinate system r, θ, ξ into r, θ', ξ', then

$$P_{n}^{\,j} (\cos \theta')e^{ij\xi'} = \sum_{k=-n}^{n} H_{jk}^{n}(R)P_{n}^{\,k} (\cos \theta)e^{ik\xi}$$

where $P_{n}^{\,j}$ $(\cos \theta)$ $e^{ij\xi}$ are the normalized spherical harmonics.

In this table the values of H_{jk}^{n} $(0, \phi, 0)$ are given for $n = \frac{3}{2}(\frac{1}{2})\frac{8}{2}$ and $\phi = 0°$ $(1°)$ $90°$.

Authors' Summary

40[I, M, P].—R. L. Murray & L. A. Mink, *Tables of $\overline{\phi^{n}}$ for Reactor Slabs, Cylinders, and Spheres, $n = 1$ to $n = 20$*, Department of Engineering Research, North

Carolina State College, Raleigh, North Carolina, Bulletin No. 70, 1958, 56 p., 28 cm. Price $1.50.

This paper contains a tabulation of average neutron fluxes for slab, cylindrical and spherical reactors. Specifically, the following functions are tabulated to 8 decimal places:

$$\text{Slab} \qquad \overline{\phi^n}(z) = \frac{1}{(\delta\pi/2)} \int_0^{\delta\pi/2} [\text{Cos } x]^n \, dx \quad \text{with} \quad x = \pi z/H$$

$$\text{Cylinder} \quad \overline{\phi^n}(r) = \frac{2}{(\delta j_0)^2} \int_0^{\delta j_0} [J_0(x)]^n x^2 \, dx \quad \text{with} \quad x = j_0 r/R_{\delta\pi}$$

$$\text{Sphere} \qquad \overline{\phi^n}(r) = \frac{3}{(\delta\pi)^3} \int_0^{\delta\pi} \left[\frac{\text{Sin } x}{x}\right]^n x^2 \, dx \quad \text{with} \quad x = \pi r/R.$$

R and H represent the extrapolated boundries of the reactor.

δ, the fraction of the total dimension over which the average is performed, is given from 0 to 1 in steps of .01, and n assumes integer values from 1 to 20. The above values of $\overline{\phi^n}$ actually represent $\overline{\phi^n}/\phi_c{}^n$, where $\phi_c{}^n$ is the central flux normalized to unity.

These tables were calculated on an IBM 650. Details of the methods used are presented, as well as a method for interpolating for values not tabulated.

The paper illustrates the application of these flux averages by means of a few worked examples.

Robert Brodsky

Department of the Navy,
Washington 25, District of Columbia

41[**P, T, Z**].—Ernest F. Johnson, "Automatic Process Control"—Chapter II, *Advances in Chemical Engineering*, Thomas B. Drew & John W. Hoopes, Jr., Editors, Academic Press Inc., New York, 1958, x + 338 p., 23 cm. Price $9.50.

This chapter should be a convenient primer for those to whom this subject is new. Terms are defined clearly, equations given and the bibliography is excellent. Some of the more advanced notions such as "three-mode control" are discussed adequately. It is unfortunate that the topic of "sampled-data systems" is no more than mentioned, as this is an area of equal importance functionally, and will be of much greater importance in the future when chemical engineers finally appreciate that digital computers are much more powerful and flexible than analog systems.

Gilbert W. King

IBM Research Center
Yorktown Heights, New York

42[**W, X, Z**].—Armour Research Foundation, *Proceedings of the Fifth Annual Computer Applications Symposium*, 1958, Sponsored by the Armour Research Foundation of Illinois Institute of Technology, x + 153 p., 23 cm. Price $3.00.

This small volume contains copies of the papers presented at the 1958 Computer Applications Symposium held on 29 October in Chicago and sponsored by the

Armour Research Foundation. A list of papers presented at this meeting is given below.

1. Operations Research and the Automation of Banking Procedures—R. A. BYERLY

2. Information Systems Modernization in the Air Materiel Command—D. E. ELLETT

3. Utilization of Computers for Information Retrieval—A. OPLER

4. Problems and Prospects of Data-processing for Defense—C. A. PHILLIPS

5. An Integrated Data-processing System with Remote Input and Output—R. D. WHISLER

6. The Role of Character-Recognition Devices in Data-processing Systems—R. L. HARRELL

7. Input-Output—Key or Bottleneck?—R. D. ELBOURN

8. Scientific Uses of a Medium-Scale Computer with Extensive Accessory Features—R. A. HAERTLE

9. The Design of Optimum Systems—R. R. BROWN

10. Computer Applications in the Numerical Control of Machine Tools—R. B. CLEGG

11. Frontiers in Computer Technology—R. W. HAMMING

12. Computer Sharing by a Group of Consulting Engineering Firms—E. M. CHASTAIN AND J. C. McCALL

13. Current Developments in Computer Programming Techniques—F. WAY, III

14. The Future of Automatic Programming—W. F. BAUER

H. P.

43[X].—VERA RILEY & SAUL I. GASS, *Linear Programming and Associated Techniques*, Bibliographic Reference Series No. 5, Published for the Operations Research Office, The Johns Hopkins University, The Johns Hopkins Press, Baltimore, 1958, x + 613 p., 23 cm. Price $6.00.

This bibliography, a revised edition of BRS-2, *Programming for Policy Decision*, March 1954, includes over 1000 abstracts of articles, books, monographs, theses, conference proceedings, etc., dealing with linear programming and related topics such as nonlinear programming, dynamic programming, and game theory. The comprehensive work embraces references on the history, progress, and application of mathematical programming. It is divided into four sections: Part I, "Introduction," covers the early development and basic concepts of linear programming; Part II, "General Theory," embraces a wide range of topics such as advanced mathematical aspects of linear programming, computational methods and machine techniques, linear inequalities and convex sets, and theory of games; Part III, "Applications," presents applications to industrial, agricultural, and military problems, and contains a basic bibliography of material related to the field of production scheduling and inventory control; Part IV, "Nonlinear and Dynamic Programming," covers the mathematical and computational aspects of nonlinear and dynamic programming. Each section is prefaced by an expository discussion of the scope and contents of the material listed.

Notwithstanding the spate of impressive publications on the applications of linear programming, its record of accomplishment does not support the glowing advocacy of it as a practical tool in management science; its contribution has been marginal at best. The lack of efficient computational techniques and machine methods precludes the use of mathematical programming as an effective vehicle in solving the multifarious and vexing problems inherent in many large-scale industrial, governmental, and military operations. The reviewer does not discount the progress already made, but wishes to stress the fact that extensive research is still required to effectively solve the complex and formidable problems of management.

In the opinion of the reviewer, the bibliography will serve as an invaluable aid and guide for obtaining a detailed account of the current mathematical techniques and applications. The broad spectrum of references describing the work in the field of linear programming makes the subject accessible even to readers without advanced training in mathematics.

Milton Siegel

Applied Mathematics Laboratory,
David Taylor Model Basin,
Washington 7, District of Columbia

44[X, Z].—R. W. Metzger, *Elementary Mathematical Programming*, John Wiley & Sons, Inc., New York, 1958, ix + 246 p., 22.5 cm. Price $5.95.

This book was written as an attempt to fill a gap which exists in the literature concerning mathematical programming. The author tries with singular success to hit the middle ground between purely literary discussions for the layman, which are of no technical value, and highly technical papers, which are incomprehensible to all except professional mathematicians. Mr. Metzger assumes his reader has a limited background in mathematics but wishes to understand the basic techniques and applications of mathematical programming.

In Chapters 2, 3, and 4 Mr. Metzger gives an excellent exposition of the distribution, simplex and approximation methods. The techniques are worked out by use of sample problems which are well chosen for their simplicity and applicability. The mathematics of the analysis and solution is complete, accurate, and understandable. Mr. Metzger goes to considerable lengths to keep the reader from getting lost in unfamiliar mathematics. At no time is the reader informed that "it is not difficult to show that . . ." Each step is carefully explained. No proofs are given for any of the methods. However, intuitive explanations keep the reader from feeling that the mathematics involved is a kind of "black box" which magically produces the right answer. Adequate references are given for those who wish to verify the proofs. Each method of solution is summarized in a step by step listing of the procedure for easy reference.

Chapter 6 is a short discussion of computers and their applications in this field. Existing programs for various computers are mentioned and some indications given as to size limitations and time required for computation.

The remaining chapters are concerned with various applications of mathematical programming to industrial and business problems. The analysis preceding and following the mathematical computation are emphasized and illustrated with somewhat simplified but practical problems.

The book is well documented, and contains a fairly good bibliography. It is best suited for individual study but could be used as a text or supplementary text in a course. The only fault lies in the scarcity of practice problems other than those worked out in full in the text.

This book will be extremely valuable to any person who wishes an introduction to mathematical programming, whether or not he is a mathematician. Mr. Metzger should be commended for fulfilling a real need.

JEAN PORTER

9308 48th Avenue,
College Park, Maryland

45[Z].—NED CHAPIN, *An Introduction to Automatic Computers*, D. Van Nostrand Company, Inc., Princeton, New Jersey, 1957, viii + 525 p., 25 cm. Price $8.75.

The stated objective of this revised edition of a 1955 book is to present computers from the business systems point of view. This is developed through two subbooks, one set of chapters dealing with the logical and engineering nature of a computer and with structural and operating characteristics of a large number of commercial devices and computers, and another set of chapters dealing with the applications of computers in business operations. The programming chapters are related to computers, not to application.

The style and depth are largely that of the numerous popularizations in business, systems and accounting periodicals. There is prevalence of many of the common clichés and generalizations, preoccupation with explaining or destroying "popular misconceptions", and some narrowness of viewpoint about the nature of business management and the role of computers in it. Many managers are still seeking conceptual simplification and education about computers and their relation to management. It is doubtful that they approach this search as less than serious students, students hoping to find scientific classification and orderly development rather than demeaning and patronizing popularization.

To its credit, this book touches on several very important aspects of computers in business—the control features of systems, operations research and management science related to system definition, organization and administration of computer activity, company planning for computers, etc. There is, however, no originality, no profundity, no sureness of touch, no indication of "battle action" with these matters. Instead, there is a hollow reflection of culled magazines. There is, for example, a dearth of case history presented from the author's vast experience, the experience which qualified him to write this book.

Because there is as yet an extremely limited literature dealing with business systems design in general and with principles of the use of computers in such systems in particular, this text should prove useful, say, at the level of undergraduate students in business administration or of the high-speed orientation courses contained in the "management development programs" of many large companies. Better to serve this purpose, there are a sizeable glossary, an extensive collection (already badly out of date) of data on commercially available computers after the manner of the ONR and BRL surveys, historical material, bibliographic references partitioned by chapter content, and even related homework assignments. The popular style should contribute to the book's palatability as an undergraduate text.

H. N. Laden

The Chesapeake and Ohio Railway Company,
Terminal Tower, Cleveland 1, Ohio

46[Z].—Louis Couffignal, *Les Notions de Base*, Gauthier-Villars, Paris, 1958, 60 p., 24 cm. Price $1.55.

The title of this little pamphlet needs to be interpreted. At the top of the title page is Information et Cybernetique, Collection Internationale.

On a fly leaf are the names of three projected items of the Collection: Boulanger, *L'Automation*; Prudhomme, *Construction des machines automatiques*; and Ducasse, *L'expression des connaissances techniques*.

In the light of these the title of this pamphlet becomes clear, namely, "the basic notions of information theory and cybernetics."

Monsieur Couffignal merely lists and comments on the basic notions; there is nothing original. In fact, a fair part of the text consists of quotations from M. Couffignal's other publications.

Wiener first published the doctrine of cybernetics in 1948. The term comes from the Greek word for pilot. Aboard ship the captain decides where to go and the helmsman works the mechanism of the ship, but the pilot between them contrives that the work of the helmsman shall achieve the object of the captain. Cybernetics is the study of means—principally feed-back—of reaching prescribed goals. Subsequently Wiener published *The Human Use of Human Beings* about the social implications of cybernetics.

Couffignal contests the scientific character of these two books. It is in the analysis of human action that Couffignal hopes to find essential principles. Man lives in an environment. His actions have an object, to affect the environment. Each action is preceded by preparation for the action, a program. Preceding the program is a decision to act, based on judgements of values. Thus a definite structure appears.

Having acted, there are three possibilities. First and simplest, the environment is affected as foreseen in the program. Or, second, the environment reacts in an unforeseeable fashion, but according to known laws. A simple example is the temperature regulation of a building; because of unknowns such as the number of times doors will open or the number and activity of the occupants, it is not possible to prescribe directly how to regulate the furnace. But a simple feed-back makes possible a program that accomplishes the desired objective. Third, the environment reacts according to unknown laws in an unforeseeable way, as a pursued butterfly.

Each of the situations above requires information about the environment; the less the reaction of the environment is understood, the more information is needed to build an effective program. Thus information is seen to be an essential of cybernetics.

The word "information" comes from the French, and originally designated the action of giving a form; it still has this connotation in its legal usage. The concept "quantity of information" is new. A specific occurrence of information has a physical form which is irrelevant to its meaning. For instance, a television program is fed over a microwave network as modulation of super-high-frequency radio waves, at a transmitting station it may be remodulated onto ultra-high-frequency or very-high-frequency waves, or recorded on magnetic tape. At a receiver it is viewed as a picture on a kinescope and it may be photographed. Each of these physical forms is different and is called a "support" by Couffignal. The common property of all the possible forms of an information he calls its "semantic."

Couffignal argues that cybernetics is not a science. Neither is it a technology nor an applied science. He concludes that it is an art, the art of insuring effectiveness of action. The material with which it works is information.

He conceives for each technique a set of systems in which the technique is effective. This set of systems is called the "domain of effectiveness" of the technique. It is intended that subsequent works in the series of which this is the first will report progress in exploring these domains. Some possibilities are: "subjective man", who reacts to his concept of his environment instead of the environment itself; "social man", who reacts to a social environment, or who reacts en mass as a social unit. The theory of governors may yet apply to Governors with a capital G. There is the domain of machines, especially information machines. A parallel domain is that of automation, the replacing of human beings by machines. There is the domain of models and simulation, which includes mathematical models (such as the Maxwell equations) of the physical world. And then there is the domain of knowledge. This is clearly basic. In fact, it is overwhelming that Couffignal stakes out such a broad claim.

The last chapter is a bald list of concepts. It is in six categories: human beings; human actions; information; mechanisms; analogues; and mentality.

Evidently M. Couffignal believes that he is starting something that will have far-reaching consequences. He has mapped out a broad outline that will take a generation to fill in. But he has only outlined broad areas with vague boundaries; subsequent items in this collection will need to be much more precise if they are to have scientific value.

H. Campaigne

National Security Agency,
Ft. George G. Meade, Maryland

47[Z].—J. von Neumann, *The Computer and the Brain*, Yale University Press, New Haven, Conn., 1958, xiv + 82 p., 21 cm. Price $3.00.

This small volume constitutes the last contribution of one of the great scientists of our time, John von Neumann. The material was prepared for delivery at Yale University during the spring of 1956 in the Silliman Lectures series. However, the lectures were never given, owing to the increasing severity of the illness which finally took von Neumann's life on February 8, 1957. The subject is one which attracted his interest for a number of years. It is a subject in which he was eminently qualified, by virtue of his great genius and his contributions in the fields of high-speed calculators and the logic of automata. In spite of the preliminary nature of this work, it is destined to become the nucleus of a new field of research which will challenge the minds of men for many years to come—the comparative study of the human brain and man-made automata.

The book is divided in two parts. In the first part von Neumann discusses the basic principles underlying the design of modern computing machines, both analog and digital. Some of the properties of analog computers receiving special attention are: (1) their continuous but approximate method for representing quantitative information; (2) their capacity to obtain valid solutions to many classes of mathematical problems in spite of their limited accuracy, which is, at best, 0.01%; and (3) the use of basic operations more advanced than the four arithmetic operations in some analog calculators such as the "Stieltjes" integral in the differential analyzer. The characteristics of modern electronic digital calculators are discussed in somewhat greater detail. Of special interest for comparison with the human brain are: (1) the quantized or "marker" method for representation of information; (2) the conventional basic arithmetic and logical operations used; (3) the organization of the controls of digital calculators, which is governed by strict logical rules and prescribed sequences; (4) the high precision available in these computers, which is approximately 10^{-11} in large-scale computers currently in operation; (5) the organization and capacity of the memory components; (6) the great arithmetic and logical "depth" required for the solution of most problems on digital computers (i.e. the very large number of operations used in sequence); (7) the use of the memory for storing instructions as well as other information, resulting in the capability of digital computers to modify or operate on instructions; and (8) the use of interpretive subroutines or "short codes" for carrying out functions more advanced than the basic operations built into the hardware of a computer. One of the key points made by von Neumann is that the very high precision requirement of digital computers is dictated by the inherent deterioration of accuracy resulting from the very large number, or great "depth," of the mathematical and logical operations used in sequence.

In the second part von Neumann compares the functioning of the human brain with the operation of a modern computer, bringing out the areas of "similarity and dissimilarity between these two kinds of automata." He begins by describing the known properties of a neuron and the apparent digital character of its mechanism for receiving and transmitting pulses. Some of the more superficial comparisons between the human brain and its man-made counterpart are: (1) *speed*—there is a factor of about 10^4 to 10^5 in favor of the man-made components; (2) *size*—there is a factor of about 10^8 to 10^9 in favor of the human brain (the number of neurons estimated in the central nervous system is of the order of 10^{10}); (3) *energy dissipation*—there is a factor of 10^8 to 10^9 in favor of the human components; and (4) *accuracy*—there is a factor about 10^8 in favor of the artificial componentry.

Looking deeper into the more intrinsic areas of comparison, von Neumann is led to the conclusion that the basic internal language used by the brain is undoubtedly quite different from the mathematical language with which we are acquainted. He arrives at this conclusion primarily on the basis of the argument that the information stored in the brain lacks sufficient accuracy to enable it to carry out mathematical and logical processes in such "depth" as would be required if the language used were based on conventional mathematical symbols. Von Neumann conjectures that the language used by the brain is probably statistical in nature in which correlation processes play an important role. (See *The Perceptron—A theory of statistical separability in cognitive systems*, Cornell Aero. Lab. Inc., Report Nos. VG 1196 G-1, 1958 and 1196 G-2, by F. Rosenblatt.) He concludes the book with the remark: "Thus logics and mathematics in the central nervous system, when received as languages, must structurally be essentially different from those languages to which our common experience refers.

"It also ought to be noted that the language here involved may well correspond to a short code in the sense described earlier, rather than to a complete code: when we talk mathematics, we may be discussing a *secondary* language, built on the *primary* language truly used by the central nervous system. Thus the outward forms of *our* mathematics are not absolutely relevant from the point of view of evaluating what the mathematical or logical language *truly* used by the central nervous system is. However, the above remarks about reliability and logical and arithmetical depth prove that whatever the system is, it cannot fail to differ considerably from what we consciously and explicitly consider as mathematics."

(Courtesy of *Applied Mechanics Reviews*) H.P.

48[Z].—Charles V. L. Smith, *Electronic Digital Computers*, McGraw-Hill, New York, 1959, xi + 443 p., 24 cm. Price $12.00.

In the preface the author states: "This is not a treatise on digital computer engineering, nor on the other hand an exhaustive treatise on logical design. The reader will find considerable discussion of circuits and components—enough, I hope, to give him a reasonably complete understanding of various ways in which the usual functions of a computer, such as memory, control, the performance of arithmetic, and input and output, can be realized physically. But I have not attempted a treatment sufficiently detailed to provide design information. The reader will also find sufficient information on computer arithmetic and instruction codes to provide him with a basic understanding of these matters, but here again I have not attempted the detailed treatment that a logical designer would demand, and I have for brevity considered only machines using binary arithmetic. My purpose has been to provide the reader with sufficient information to understand how digital computers function."

No book of this size could possibly cover every phase of computers; however, for the numerous topics chosen, the author's objective has been met. Mathematicians, programmers, and engineers who have limited knowledge of the basic principles should find this volume especially useful in extending their understandings. As stated previously, this is not a treatise, and its organization is such that it is probably most useful as a reference text. There are many additional references footnoted throughout the book for those desiring additional details. The usefulness of the subject index, however, suffers somewhat because of its brevity. The bulk of the material has been written about computers and techniques of 1956 and earlier, but this need not detract in any way from the value of the material.

Chapter titles are:

1. Digital-computer Arithmetic (Number Systems and Their Machine Representation)

2. Instruction Codes

3. Some General Considerations on Systems (Basic Functions and Inner Structure)

4. Basic Logic Circuits and Their Representation (Logic or Switching Circuits)

5. Static Memory Cells (Circuits and Devices Having Two Stable States)

6. Dynamic Memory Cells (Circuits and Devices Which Require Recirculation or Regeneration)

7. Higher-order Logic Circuits (Boolean Algebra, Switching Matrices)

8. Shifting Registers (Basic Storage Combined into Complex Structures)

9. Counters

10. Adders and Accumulators

11. Large-scale Memory Devices I (Magnetic Drums, Ultrasonic and Magnetostrictive Delay Lines)

12. Large-scale Memory Devices II (Electrostatic Tubes and Magnetic Cores)

13. The Arithmetic Unit

14. The Memory

15. The Central Control

16. Input and Output

17. Superspeed Computers (NBS, LARC, STRETCH, ILLIAC II).

GORDON D. GOLDSTEIN

Office of Naval Research,
Washington, District of Columbia

TABLE ERRATA

272.—H. M. Nautical Almanac Office, *Planetary Coordinates*, 1960–1980, H. M. Stationery Office, London, 1958.

P. 145, line 18; *for* n^3 *read* n^2.

P. 146, line 14; *for* 59.8 *read* 58.1.

This erroneous value for m is also on page xiv; the correct value brings the results of the example into still closer agreement with those on page xiii.

J. G. Porter

H. M. Nautical Almanac Office,
Royal Greenwich Observatory,
Hailsham, Sussex, England

NOTES

Vychislitel'na͡ia Matematika

The new Russian Journal *Vychislitel'na͡ia Matematika* has recently appeared. The subjects treated in the first two issues are in the fields of Numerical Methods and Computing Machines (both digital and analogue). We reprint below the table of contents of the first two volumes.

1. Izdatel'stvo Akademii Nauk SSSR, Vychislitel'nyĭ T͡sentr, Sbornik 1, Moskva, 1957.

L. A. L͡iusternik; The finite-difference analog of Green's function in the three-dimensional case.

I͡U. V. Vorob'ev; The method of moments in the problem of the vibrations of linear systems.

E. A. Volkov; Investigation of a method of improving the accuracy of the method of nets for solving Poisson's equation.

V. K. Saul'ev; On a class of elliptic equations, solved by use of the method of finite differences.

V. K. Saul'ev; On an estimate of the error in getting characteristic functions by the method of finite differences.

A. I. Vzorova; On the construction of polynomials orthogonal on a family of ellipses.

I͡A. I. Alikhashkin; A method of calculating the discharge for forced flows through an imperfect slit.

I͡A. I. Alikhashkin; Solution of the problem of the imperfect slit by the method of straight lines.

G. S. Khovanskiĭ; On the replacement of the logarithm function by a power in making approximate nomograms.

S. P. Kapi͡tsa; Mechanical computation of the harmonic adjoint function.

2. Izdatel'stvo Akademii Nauk SSSR, Vychislitel'nyĭ T͡sentr, Sbornik 2, Moskva, 1957.

M. R. Shura-Bura; Approximating functions of many variables by functions each of which depends on one variable.

P. I. Chushkin; The flow around ellipses and ellipsoids in sonic gas flow.

O. N. Ka͡tskova and I͡U. D. Shmyglevskiĭ; Axisymmetric supersonic flow of a freely expanding gas with a plane transitional surface.

V. S. Linskiĭ; Computation of elementary function on automatic computing machines.

M. G. Rappoport; Computation of finite differences on punched-card machines.

B. M. Drozdov and M. G. Rappoport; Coding of operations on the electronic computer EV80-3.

A. G. Gesse; Electrical resonance networks for solving systems of linear equations.

G. S. Khovanskiĭ; Methodology of constructing nomograms with a triangular (hexagonal) transparency.

E.I.

New Journals

The publication of two new German journals in the fields of numerical analysis and data-processing has been announced recently. The journal in numerical analysis has the title *Numerische Mathematik* and is published by Lange & Springer, West-Berlin. The editors are: R. Sauer, E. Stiefel, J. Todd, and A. Walther. The first issue contains the following papers:

1. Über diskrete und lineare Tschebyscheff-Approximationen.
2. On certain methods for expanding the characteristic polynomial.
3. Orthogonal polynomials in several variables.
4. Report on the algorithmic language ALGOL.

The new journal on data-processing is published by Friedr. Vieweg & Sohn, Braunschweig under the editorship of H. K. Schuff. Its title is *Elektronische Datenverarbeitung*. The first issue contains a number of papers on programming and application of digital computers with particular emphasis on their use in the management data analysis field.

H.P.

ysis
ysis
ger,
The

ohn,
ten-
ap-
age-

P.